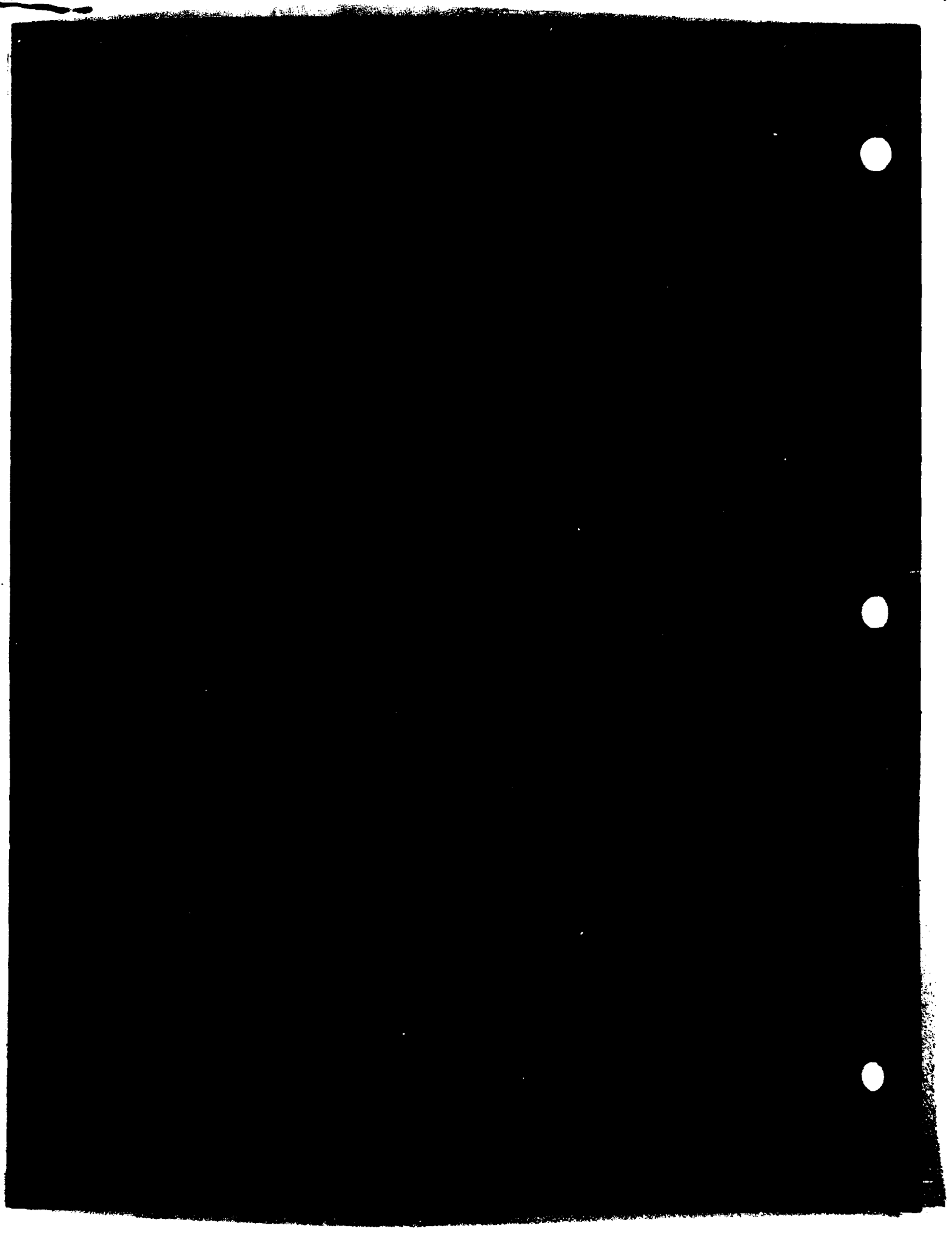
File Copy

AD-A210 257



DISCLAIMER NOTICE

THIS DOCUMENT IS BEST QUALITY PRACTICABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION: Unclassified	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S): Dept of Mathematical Sciences	5. MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION: Dept of Mathematical Sciences; 6b. OFFICE SYMBOL (If applicable): DFMS	7a. NAME OF MONITORING ORGANIZATION
6c. ADDRESS (City, State and ZIP Code): US Air Force Academy, CO 80840-5701	7b. ADDRESS (City, State and ZIP Code)
8a. NAME OF FUNDING/SPONSORING ORGANIZATION; 8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
8c. ADDRESS (City, State and ZIP Code)	10. SOURCE OF FUNDING NOS.

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT NO.

11. TITLE (Include Security Classification): An Introduction to Fractals and Chaos

12. PERSONAL AUTHOR(S): John Haussermann, LT, USN

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Yr., Mo., Day)	15. PAGE COUNT
Final	FROM Jun 88 TO Jun 89		

16. SUPPLEMENTARY NOTATION

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB. GR.	

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

Fractals and chaos have gained immense popularity in the last few years. Fractal objects and chaotic dynamics are being postulated and observed in diverse areas of science. This report gives a mathematical introduction to fractals and chaos, including some necessary background for the technical definitions.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT	21. ABSTRACT SECURITY CLASSIFICATION
UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS ☐	

22a. NAME OF RESPONSIBLE INDIVIDUAL	22b. TELEPHONE NUMBER (Include Area Code)	22c. OFFICE SYMBOL
John Haussermann, LT, USN	(719) 472-4470	DFMS

DD FORM 1473, 83 APR EDITION OF 1 JAN 73 IS OBSOLETE.

SECURITY CLASSIFICATION OF THIS PAGE

AN INTRODUCTION TO FRACTALS AND CHAOS

by

John W. Haussermann, LT, USN
Department of Mathematical Sciences

TABLE OF CONTENTS

Appendix 1: "Tomorrow's Shapes: The Practical Fractal", THE ECONOMIST, 26 Dec 87, pages 97-101.

Appendix 2: "Fractal Applications", Mort La Brecque, MOSAIC, Winter 86/87, Vol. 17 No. 4, pages 34-48.

Appendix 3: "Classical Chaos", Roderick V. Jensen, American Scientist, March/April 1987, Vol. 75, pages 168-187.

Appendix 4: "Chaos, Strange Attractors, and Fractal Basin Boundaries in Nonlinear Dynamics", Celso Grebogi, Edward Ott, James A. Yorke, SCIENCE, 30 Oct 87, Vol. 238, pages 631-638.

Appendix 5: "A Better Way to Compress Images," Michael Barnsley and Alan Sloan, BYTE, January 1988.

Acknowledgements

I would like to thank: Colonel Daniel W Litwhiler and the DFMS Personnel Council for allocating an extra Math 495 slot during spring, 1989; Lt Colonel Stephen R. Schmidt for giving me the time to do this research; and the cadets whose interest generated the format of this report and the "Special Topics" course to be taught during spring, 1989.

I would also like to thank the publishers concerned for permission to reprint the articles appearing in appendices 1 through 5. These reprints are intended solely for use in this report and should not be duplicated without permission from the applicable journal.

Mrs Alice Wilmoth put many hours of work into this report. Her effort made these pages a reality--in spite of equipment failure and haphazard changes by myself. I thank her.

Finally, the arduous task of editing this report was greatly simplified by the time and effort of the technical reviews: Lt Col James Aubry, Major Mark Rogers, Major Robert Sheldon, and Captains Kirk Yost and Keith Bergeron. Their comments and criticisms were invaluable for improving the continuity and accuracy of this work.

PREFACE

Recent popular literature abounds with articles on fractals or chaos. Laymen will find them quite accessible, except for occasional stumbling blocks like dimension, self-similarity, dynamics, randomness, etc. But technical readers will likely be frustrated by a lack of detail. This report assumes the reader has a mathematical background equivalent to an upper-division math major. The report is being used as the textbook for a Math 495 course during spring, 1989.

The popularity of fractals and chaos is not an accident. There is an amazing number of natural phenomena which can be interpreted by either fractals or chaos or both. It is more surprising that these fields weren't "discovered" until the mid-1970's. Benoit Mandelbrot noticed that many geographic entities--coast lines, mountains, etc.--could be thought of as having a "fractional dimension"; hence, fractal. Mitchell Feigenbaum theorized universal properties of certain "chaotic processes". For example, fluid turbulence and a nonlinear oscillator exhibit similar qualitative and quantitative behavior.

Mathematicians became interested when old topics like Hausdorff Dimension, Ergodic Theory, and everything in between were found to have applications and to have the ability to make pretty computer graphics. Fractals and chaos have become such hot topics that they have worked their way into the diverse fields of art and biology.

The eager researcher will jump at (or contrive) any opportunity to say "Ha, a fractal!" or, "Ha, chaos!" But: FRACTALS AND CHAOS ARE NOT MODELS OF NATURE. THEY ARE, AT BEST, SYMPTOMS. It is like noting that the sky is blue.

Quite often, knowing an object is a fractal or a process is chaotic will not give any new information about that object or process.

In the author's opinion, the classification of something as a fractal or chaotic process is most useful from a "first principles" point of view. For example, it has been known for some time that clouds are self-similar (fractals) on a scale of about 10 km on down. More recently, from satellite photographs, it was found that they're actually self-similar on a scale of 1000 km on down! This information is useful in the sense that any valid theory on the formation of clouds must address this large-scale self-similiarity. Likewise, population growth in nature exhibits chaotic behavior. So, there are definite limits to a non-chaotic (exponential growth, for instance) model's applicability.

The goal of this report is to give a basic foundation upon which the interested researcher may "build-to-suit". To this end, and for the Math 495 students, exercises will be found at the end of many sections. There are several computer programs available (for the Z-248 with EGA) which illustrate the ideas presented in this report.

If this is your first encounter with the ideas in and the applications of fractals and chaos, I hope you find them as enjoyable and amazing as I do.

Reading List

In addition to the articles reprinted in appendices 1 through 5, it is highly recommended that the following books be read by the interested researcher (or layman):

1. CHAOS: Making a New Science, James Gleick, 1987, Viking Penguin Inc., ISBN 0-670-81178-5.

Comments: This is the best book on science directed to the layman that I have read. It is also the best book on chaos that I have read. The author does an outstanding job giving historical perspective, the insights of leaders in the field, and relevance to today's society.

2. The Fractal Geometry of Nature, Benoit Mandelbrot, 1983, W. H. Freeman and Company, ISBN 0-7167-1186-9.

Comments: In my opinion, this book is written in a egocentric style--a lot of first person usage. It also vacillates between the technical and obscure, and the obvious. The book is hard to read. However, there are many beautiful and worthwhile ideas spread throughout and punctuated by fantastic images.

3. Fractals Everywhere, Michael Barnsley, 1988, Academic Press/Harcourt Brace Jovanovitch, ISBN

Comments: This is a textbook on Iterated Function Systems (IFS).

LIST OF FIGURES

All of the figures in this report were created with programs written in Turbo Pascal 4.0 on a Zenith 248 with EGA. They are in the public domain.

A figure was printed by first getting it graphed on the monitor and then using RSDLASER.COM tc screen dump it into a file. This file was then output on a laser printer using the USAFA Local Area Network.

Part I: Fractals

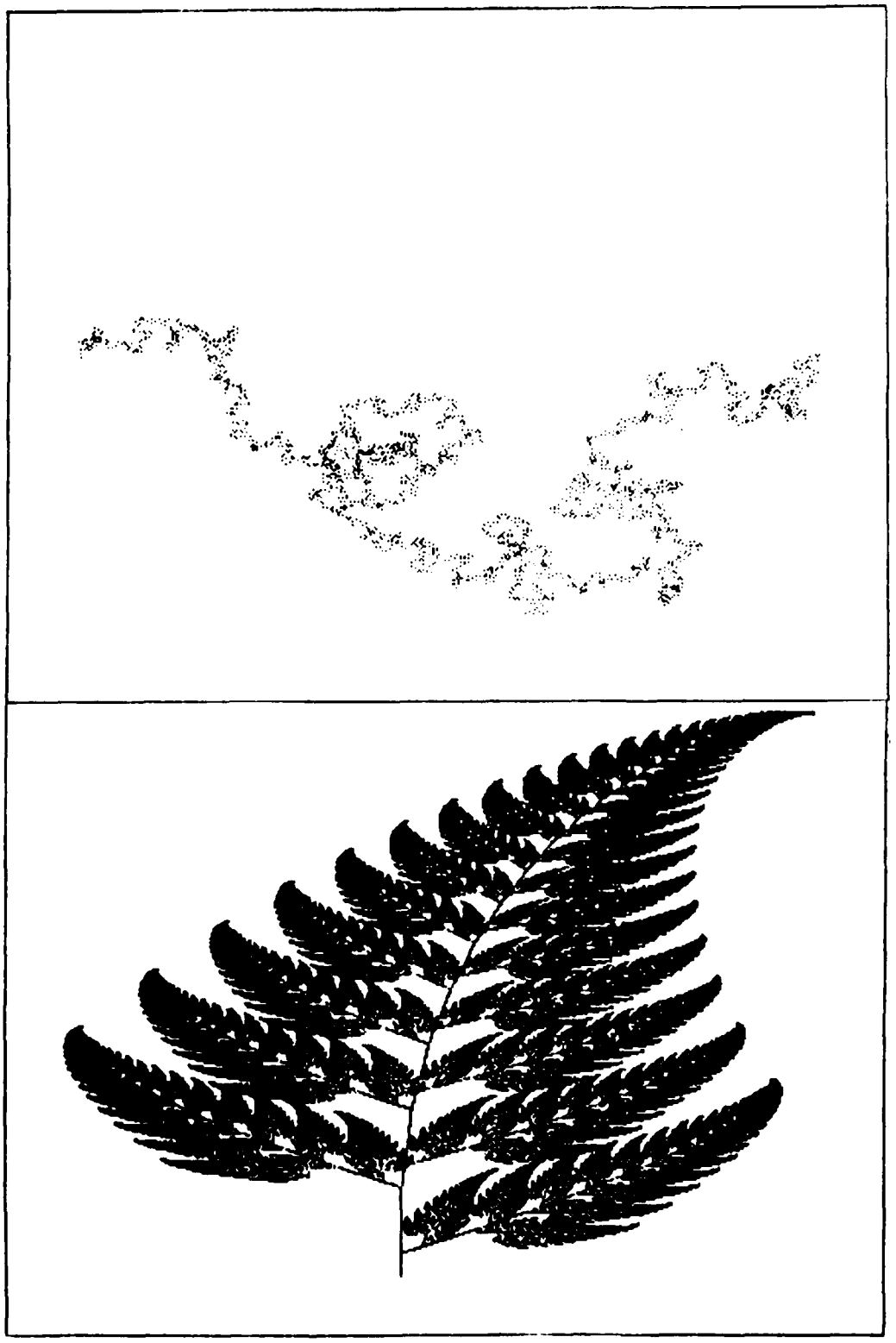


Figure 1

I.1 Introduction

It is hard to date ideas to their true originators. But "fractional dimension" is largely due to the work of Felix Hausdorff (1868-1942). This was a mathematical formulation and was not necessarily thought to have real-world applications.

Scientists and philosphers have noticed basic patterns and "self-similarity" in nature throughout history. However, it is probably safe to say that Benoit Mandelbrot was the first person to successfully integrate fractional dimension with self-similar natural phenomena. He was the one to coin the term "fractal"--without explicitly defining it at first, but later giving it a mathematical definition.

Much of the popularity of fractals is a result of the relatively recent advances in computer graphics. Mandelbrot, who works for IBM, was able to turn out some spectacular images of fractals simulating geographic shapes--like a false Earth-rise as seen from a false moon. Many scientists and computer enthusiasts have expanded on his work. Fractal images now seem to be the cover picture of choice on calculus textbooks.

Some of the most promising work in fractals has been done by Michael Barnsley of Georgia Tech. (He is funded by DARPA and AFOSR, among others.) Barnsely's work centers on the question: Given a digitized image on a monitor, (with resolution of, say, 1024 by 1024 pixels), can fractal techniques be used to reduce the amount of memory required to store the image? The compression ratio is the quotient of the raw data of the image on the screen to the data required when using some fractal technique. The objective is to make the compression ratio as large as possible.

Current techniques (nonfractal) are able to achieve compression ratios as high as 10 to 1. Barnsley has achieved ratios of 15 to 1 for perfect replication of the image and as high as 1000 to 1 for non-perfect replications. Even though the latter are non-perfect, they retain the fundamental qualities of the picture, and look rather like a painted (or sketched) copy of a photograph.

Barnsley has formed his own company to exploit his techniques: Iterated Systems, Inc., named after the process used to store and regenerate these images. Iterated Function Systems (IFS), as a mathematical topic, will be discussed in this report.

I.2 Intuitive Dimension

What is dimension? Mathematics students might say that it's the number of basis vectors in a vector space. This is a good answer if the question is: What is the dimension of a vector space?

A vector space is an abstract structure which happens to be used a great deal to model different settings of physical phenomena. For example, we usually imagine ourselves as moving around in R^3, Euclidean 3-space. Relativity is modeled using a curved 4-dimensional vector space called space-time. But vector spaces are used almost exclusively to model some section of the universe, rarely for the objects within.

Finding a dimension is intuitively like counting "degrees of freedom." How many (seemingly) independent observations are possible? We might say that colored pigment is 3-dimensional since it can be decomposed into red, green, and blue. Sound is infinite dimensional, but hearing probably isn't, since humans cannot hear the whole auditory spectrum. Taste might be 5-dimensional.

These examples deal with information. How much information is the key question in determining dimensions this way. However, there is a difference between information and "shape". We will define a shape to be what's left when all the available information is known. Most objects are a combination of shape and information.

Basis vectors are part of the "where" in a shape. They are basis vectors because they are all that is needed to convey this information. So counting basis vectors is equivalent to finding an information dimension.

We will use historical techniques to find dimensions of objects which aren't vector spaces. It should come as a relief that shape and information dimensions will agree when they can be applied to the same object, like R^3. It is unfortunate, however, that only information dimension seems to be intuitive. The definitions we will use for "shape" dimension are difficult.

1.3 Observations and Representations

Before we develop methods of determining the dimension of shapes, we must distinguish between different dimensions used to represent an object. We will use 3-dimensional Euclidean Space, R^3, to model the three directions which seem to surround our environment. Two-dimensional Euclidean space, R^2, will be used to model a sheet of paper or a computer screen.

The subjective link in any observation is the human. Our vision is influenced by imperfect eyes and an imperfect brain. But essentially, reflected (or transmitted) light from an object in R^3 is imaged (through the iris and cornea) onto the retina at the back of the eye. Information is sent from the retina via the optic nerve to the brain. Because the retina is a surface (and can be modeled with 2 dimensions), there are subtle ways in which the brain interprets a 3-dimensional image from the information sent by the eye. The study of this process is called stereopsis.

Perhaps because of the way our vision works, humans are able to interpret 3-dimensional images from photographs or drawings in R^2. And because it is easier to take a picture than build a model (or construct a hologram), we commonly use R^2 to represent images from R^3.

Definition: A set, S, is imbedded in another set, T, if and only if $S \subseteq T$.

For example, if we take a picture of a ball (imbedded in R^3), the photograph will contain a shaded disc imbedded in R^2. The area of the shaded disc is a two-dimensional subset of R^2. The shading cues our eyes to see curvature which is not present in the photograph.

Definition: We will say an n-dimensional object, E, is flat, if E can be imbedded in R^n; i.e., $E \subseteq R^n$.

Thus, a line is flat (straight), since it is one dimensional and can be imagined to lie in R (the number line)--it is the whole set in this case. A shaded disc, (filled in circle), is flat since it is two-dimensional but fits in R^2. A solid box is flat, since it is three-dimensional and a subset of R^3.

A triangle is not flat. Neither is a circle. Both curves cannot be imbedded in R even though they are 1-dimensional. These curves can be imbedded in R^2. However, a helix (a curve in the shape of a spring) cannot even be imbedded in R^2. It is still 1-dimensional, but must be imbedded in R^3. Fortunately, all 1-dimensional objects (curves) may be imbedded in R^3.

A surface is a 2-dimensional object. A sphere (the surface of a ball) is curved but can be imbedded in R^3. An eggshell is also not flat. (A sphere cannot be imbedded in R^2.) But there are surfaces, like the Klein Bottle, which cannot be imbedded in R^3 and must be imbedded in R^4. The situation with "curved" space (3-dimensional objects) is even worse. By definition, we cannot "see" any curved space since we can only see objects imbedded in R^3. One can imagine that R^4 is not even enough to imbed all curved 3-dimensional spaces, so the space-time model (which is curved 4-dimensional space) must live in a very high dimensional R^n.

The point of this discussion is to be able to recognize the dimension of the object as being (possibly) different from the space in which it is imbedded, and to realize that visual cues (or tricks) are necessary to represent curved 2-dimensional objects (and some 1-dimensional objects) in R^2, (a picture).

Exercises:

1. What is the dimension of a pencil? Is it flat? What is the (approximate) dimension of the surface of a pencil? Is it flat?

2. A torus can be constructed by taking a finite circular cylinder and joining the two ends to make a shape like an inner tube. What is the dimension of the torus? What is the smallest dimensional Euclidean space into which it can be imbedded?

3. A Klein Bottle is constructed like a torus, except the two ends are not joined in the natural way. Instead, they are attached so that the cylinder is on the same side of the join. Is this possible in R^3? What is the dimension of the Klein Bottle?

4. Suppose you take a picture (of the surface) of a pyramid. In the photograph, what dimension is the surface of the pyramid? Is it flat? What is the dimension of the surface of the actual pyramid? Is it flat?

I.4 Topology and Open Covers

The objective of part I is to introduce the reader to fractional dimension. But first, we need to introduce "topological dimension" as a base. The topological dimenision will not be fractional, but some ideas from its development are extrapolated in the formulation of Hausdorff dimension, which is fractional.

Topology is the study of shape without distance. Because it is so abstract, many shapes which we perceive as different are lumped together by topology. For example, the surface of a cube, a sphere, and the surface of a football are all the same. Similarly, a semi-circle and a line segment are equated in topology.

So, most of the aspects of curvature are ignored by topology. For this reason, a topological definition of dimension is very useful. To use topology, it is not necessary to throw away our familiar mathematical structures (distance, arithmetic, etc.). The ideas of topology are just incorporated into one's setting. These ideas are formulated using open sets:

Definition: Let T be a set and Ω a collection of subsets of T. (T, Ω) is called a topological space (with topology Ω) if and only if:

(i) $\phi \in \Omega$ and $T \in \Omega$; (ϕ is the empty set).

(ii) If $G_\alpha \in \Omega$ for each $\alpha \in I$ (I is any index set), then $\bigcup_{\alpha \in I} G_\alpha \in \Omega$.

(iii) If $G_i \in \Omega$ for $i = 1, \ldots, n$, then $\bigcap_{i=1}^{n} G_i \in \Omega$.

A topological space, T, is a set together with a topology on that set. Elements of the topology are subsets of T. They are called open sets. To be a topology, three criteria must be satisfied: (i) the empty set and T itself must be open; (ii) any union of open sets must still be open; and (iii) any finite intersection of open sets must still be open.

We will work in familiar spaces, like R^2 and R^3, and define a topology on them. This will make topological dimensions accessible in our settings.

Examples:

1. (R, Ω_1) is a topological space where R is the real line and a subset of R is in Ω_1 (the topology) if and only if it can be written as a (possibly) infinite union of open intervals. We throw in the empty set to be complete.

Thus, each open interval (including R) is itself an open set. Because two open intervals must intersect in an open interval or an empty set (both of which are open), criterion iii is satisfied. Criterion ii is satisfied by definition.

The open intervals are called a basis for the topology since any open set is a union of these basis sets. (They act like components.)

2. (R^2, Ω_2) is a topological space where Ω_2 is formed from the basis consisting of all open discs in R^2. (An open disc is the area inside a circle, without the circle itself.) Therefore, open sets in Ω_2 are unions of open discs. It is a little harder to verify that finite intersections of open sets are still open.

3. (R^3, Ω_3) is a topological space when Ω_3 consists of open sets formed from the basis of open balls (the interiors of spheres).

Remark: In examples 2 and 3 above, it is possible to change the basis without changing the topology. See the exercises.

Definition: A set, F, in a topological space, T, is called closed if and only if the complement of F, denoted F^c, is open.

Theorem: If (T, Ω) is a topological space then:

(i) Both ϕ and T are closed;

(ii) if F_α is closed for each $\alpha \in I$ (I is any index set), then $\bigcap_{\alpha \in I} F_\alpha$ is closed and

(iii) if F_i is closed for $i = 1, \ldots, n$, then $\bigcup_{i=1}^{n} F_i$ is closed.

The proof of the theorem relies on the definition of closed as the (set) complement of open and is left for the exercises.

When one is working in a metric space, like R, R^2, or R^3, there is a natural topology induced by the metric. The topologies in examples 1 through 3 are induced by the Euclidean metrics. For this reason, limit points and interior points are other ways to characterize closed and open sets.

Definitions: Let (R^n, d) be Euclidean n-space with the usual metric, d:

$$d[(x_1, \ldots, x_n), (y_1, \ldots, y_n)] = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} .$$

Let $A \subseteq R^n$ and $x_0 \in R^n$. Then x_0 is an interior point of A if and only if there is an $\varepsilon > 0$ so that if $d(x, x_0) < \varepsilon$ then $x \in A$. Also, x_0 is a limit point of A if and only if for every $\varepsilon > 0$ there is an $x \neq x_0$ so that $x \in A$ and $d(x, x_0) < \varepsilon$.

Theorem: In examples 1 through 3, a set, A, is open if and only if each element of A is an interior point of A. Also, a set, B, is closed if and only if B contains all of its limit points.

The proof is left for the exercises.

In topology, open sets have a connotation of being large. (In a metric space, every point of an open set is an interior point; so open sets cover a lot of area.) One way to measure the size of an arbitrary set, is to find the "smallest" open set which contains it.

Definition: Let (T, Ω) be a topological space, and suppose $A \subseteq T$. Then U is an open cover of A if and only if $U \in \Omega$ and $A \subseteq U$.

There are usually (infinitely) many open covers for the same set. In subsequent sections, we will find ways of measuring the "smallest" such cover. This is part of the subject called measure theory.

Example 4.

There is no "smallest" open cover of a single point, $x_0 \in R$, since any interval of the form $(x_0 - \varepsilon, x_0 + \varepsilon)$ is an open cover of $\{x_0\}$ $(\varepsilon > 0)$.

But, the "smallest" open cover of any open set is that set itself.

We will end this section with another definition of "large".

Definition. In a topological space (T, Ω), a subset $A \subseteq T$ is dense in T if and only if for each U in Ω, $U \cap A \neq \phi$.

Exercises:

1. Let X be a nonempty set and P(X) be the power set of X. So P(X) contains all subsets of X. Show that $(X, P(X))$ is a topological space. Which subsets of X are closed?

2. In R, let $(a, b] = \{x \in R: a < x \leq b\}$ be a half-open interval. Define a new topology on R by: a set is open if and only if it can be written as the union of half-open intervals (always closed on the right), or is the empty set. Show that an open interval, $(u, v) = \{x: u < x < v\}$, is an open set in this new topology. Thus, the new topology contains the usual topology. Is $(a, b]$ open in the usual topology? The new topology is finer than the usual topology; or the usual topology is coarser than the new topology.

3. Using the definition of closed in a topological space, (T, Ω)...

   a. Show that ϕ and T are closed.

   b. Show that if F_α is closed for $\alpha \in I$, then $\bigcap_{\alpha \in I} F_\alpha \in \Omega$.

   c. Show that if F_i is closed for $i = 1, \ldots, n$, then $\bigcup_{i=1}^{n} F_i$ is closed.

Hint: $\left(\bigcup_{\alpha \in I} A_\alpha\right)^c = \bigcap_{\alpha \in I} A_\alpha^c$ and $\left(\bigcap_{\alpha \in I} A_\alpha\right)^c = \bigcup_{\alpha \in I} A_\alpha^c$.

4. The Euclidean metric on R is given by absolute value:

$$d(x, y) = |x - y|.$$

Show that every point of an open interval is an interior point of that interval; thus, open sets are the same metrically and topologically (with the usual topology).

5. Same as 4. with R^2.

6. Same as 4. with R^3.

7. Show that if a set doesn't contain all of its limit points, then it's not closed by showing that the missing limit point cannot be an interior point of the complement of the set. Hence, the complement of the set isn't open. Work in R^n with the Euclidean metric.

8. Show that if a set isn't closed (so that it's complement isn't open) then there is a point in the complement which isn't an interior point of the compliment, but must be a limit point of the set. Hence, the set doesn't contain all of its limit points. Work in R^n.

9. It is often possible to find a different basis for a topological space which will give the same topology:

a. Show that the basis of open squares (the interior only) in R^2 will also generate the usual topology. Is there a metric which corresponds to this basis?

b. Show that the basis of open cubes (the interior only) in R^3 will also generate the usual topology.

10. Show that a single point is a closed set in R^n. Hence, any finite subset of R^n is also closed.

11. Let Q denote the rational numbers. Is Q a closed subset of R? Is it open?

12. In R^n with the Euclidean metric and usual induced topology, open n-balls can be used as the basis of the topology. (An open 1-ball is an open interval, etc.) Show that an open set (which is a union of open n-balls) cannot be the uncountable union of <u>disjoint</u> open n-balls. (Hint: Use the fact that there are countably many points whose coordinates are all rational.)

13. Let $\{r_n\}$ be an enumeration of all the rational numbers in R. Find an open cover of $\{r_n\}$ so that the sum of the lengths of all the component intervals is less than 1.

(Hint: Let each r_n be the midpoint of an interval of length $\frac{1}{2^n}$.)

Thus, this open cover is "large" in two senses: it's open and dense. But it's "small" in the sense that it has a short length compared to all of R.

I.5 Topological Dimension

We will be interested in calculating the dimension of objects in R^2 and R^3. An important fact is that subsets of a topological space are also topological spaces with the "relative topology."

Definition: Let (T, Ω) be a topological space, and $S \subseteq T$. Define the relative topology on S to be $\Omega' = \{U' : \text{there is a } U \in \Omega \text{ with } U' = S \cap U\}$.

It is easy to show that (S, Ω') is a topological space. The open sets in S are just intersections of open sets in T with S itself.

Example 1:

Find the relative topology on a circle imbedded in R^2.

Since an open set in R^2 is a union of open discs, the relative topology on S consists of open sets which are the union of open arcs.

The following definition is due to Henri Lebesque (1875-1941). We have adapted it to subsets of R^n.

Definition: Let X be imbedded in R^n, for some n. The topological dimension (or Lebesque dimension or covering dimension) of X is less than or equal to m, i.e., $\dim X \leq m$, if and only if for any finite collection of p open sets, G_i, (relatively open in X) with $X = \bigcup_{i=1}^{p} G_i$, there is another collection of p open sets, H_i, so that each $H_i \subseteq G_i$, $X = \bigcup_{i=1}^{p} H_i$, and any $m + 2$ of the H_i have no point in common. And, if $\dim X > m - 1$, then $\dim X = m$.

Essentially, the above definition finds the most efficient open covers (in terms of overlap) of a set, X. The topological dimension is related to that efficiency.

Example 2:

A single point has topological dimension zero.

Suppose $x_0 \in R$. The only open sets in $\{x_0\}$ are ϕ and $\{x_0\}$ itself. So if G_i, $i = 1, \ldots, p$, are any open sets so that $\{x_0\} = \bigcup_{i=1}^{p} G_i$, then at least one of the G_i must be $\{x_0\}$ itself, call it (without loss of generality), G_1. Then define $H_1 = G_1$, and $H_2 = \ldots = H_p = \phi$. Thus, $\{x_0\} = \bigcup_{i=1}^{p} H_i$, but any two of the H_i have an empty intersection. Therefore, dim $\{x_0\} \leq 0$. Also, the only set with dimension -1 is the empty set. (Since each open set in the cover must be empty.) So we can conclude that dim $\{x_0\} = 0$.

Remarks: In our definition of topological dimension, we do not specify which R^n the set X is imbedded in. In example 2 we imbedded a point in R, but we could have imbedded it in R^2, R^3, etc. The topological dimension will be invariant (with respect to the imbedding space) since the usual topology of R^m is also the relative topology on R^m if R^m is imbedded in R^n where $n > m$. See the exercises.

Example 3.

The dimension of any line segment (finite or infinite) is one.

Let S be a line segment. We imbed S as an interval (open, closed, or half-open) in R. The proof will proceed by induction on the number of open sets (in S) which cover S:

If $S = \bigcup_{i=1}^{1} G_i$ then we can choose $H_1 = G_1$ and we're through.

Assume that when $S = \bigcup_{i=1}^{p} G_i$ there exists open $H_i \subseteq G_i$ so that $S = \bigcup_{i=1}^{p} H_i$ and no 3 of the H_i have a point in common.

Now suppose $S = \bigcup_{i=1}^{p+1} G_i'$ where $p \geq 1$. For $i = 1, \ldots, p-1$ define $G_i = G_i'$. We proceed through four cases based on the relationship of G_p' to G_{p+1}':

Case 1: $G'_p \cap G'_{p+1} = \phi$.

Then let $G_p = G'_p \cup G'_{p+1}$. Now, $S = \bigcup_{i=1}^{p} G_i$ so there are open $H_i \subseteq G_i$ so that $S = \bigcup_{i=1}^{p} H_i$ and no 3 of the H_i have a point in common. Define $H'_i = H_i$ for $i = 1$, ..., $p - 1$ and let $H'_p = G'_p \cap H_p$ and $H'_{p+1} = G'_{p+1} \cap H_p$. Since $G'_p \cap G'_{p+1} = \phi$, we also have $H'_p \cap H'_{p+1} = \phi$. So each H'_i is open, $H'_i \subseteq G'_i$, $S = \bigcup_{i=1}^{p+1} H'_i$, and no 3 of the H'_i have a nonempty intersection, completing the proof in this case.

Case 2: $G'_{p+1} \subseteq G'_p$.

As above, let $G_p = G'_p \cup G'_{p+1} = G'_p$. Then, we can obtain $H_i \subseteq G_i$. Redefine $H'_i = H_i$ for $i = 1$, ..., p and let $H'_{p+1} = \phi$. As before, the open $H'_i \subseteq G'_i$, $S = \bigcup_{i=1}^{p+1} H'_i$, and no 3 of the H'_i have a nonempty intersection (since the H_i's didn't), completing the proof in this case.

Case 3: $G'_p \subseteq G'_{p+1}$.

This case is the same as case 2 with p replaced by p + 1 and visa versa.

Case 4: Not case 2 or 3, but $G'_p \cap G'_{p+1} \neq \phi$.

Again, we let $G_p = G'_p \cup G'_{p+1}$ and obtain open H_1, ..., H_p. Now, let $H'_i = H_i$ for $i = 1$, ..., $p - 1$ and define $H''_p = H_p \cap G'_p$. Also, define $H''_{p+1} = H_p \cap G'_{p+1}$. Since H''_{p+1} is a union of disjoint open intervals, we will remove any interval from H''_{p+1} which is a subset of H''_p and any interval from H''_p which is a subset of H''_{p+1}, forming new sets, H'_{p+1} and H'_p, which are still open. We now have open $H'_i \subseteq G'_i$ so that $S = \bigcup_{i=1}^{p+1} H'_i$. Since no three of the H_i have a nonempty intersection, if 3 of the H'_i have a point in common, then H'_p and H'_{p+1} must be two of those sets. (Also, no 4 of the H'_i can have a nonempty intersection.) Suppose $1 \leq k < p$ and $H'_k \cap H'_p \cap H'_{p+1} \neq \phi$. Then, this nonempty intersection must be mimicked among

their open (component) interval(s). Let $I_i \subseteq H_i'$ be such component intervals for $i = k$, p, and $p + 1$. Because of the way H'_{p+1} was constructed, $I_p \not\subseteq I_{p+1}$ and $I_{p+1} \not\subseteq I_p$. We have three cases:

Case a: $I_p \cup I_{p+1} \subseteq I_k$. In this case, we may redefine H'_p and H'_{p+1} to be $H'_p - I_p$ and $H'_{p+1} - I_{p+1}$ respectively. The sets are still open, and the union of the H_i''s still cover S.

Case b: $I_k \subseteq I_p \cup I_{p+1}$. In this case we can redefine H'_k as $H'_k - I_k$.

Case c: Not case b or c. Then since $I_p \cap I_{p+1} \neq \phi$. $I_p \cup I_{p+1}$ is an interval and I_k must overlap either the right or left side. Redefine I_k so that it still overlaps but does not go so far as to also overlap $I_p \cap I_{p+1}$. Then the new H'_k will use this new I_k.

Case a, b, and c may be applied to any of the component intervals which behave in that way. After such applications, we obtain open $H_i' \subseteq C_i'$ so that $S = \bigcup_{i=1}^{p+1} H_i'$ and no 3 of the H_i' have a nonempty intersection. This completes the proof by induction and establishes that dim $S \leq 1$. Suppose S is covered by 2 overlapping open intervals, G_1 and G_2. Then, in order to still cover S, H_1 and H_2 must also be overlapping. Thus, dim $S \not\leq 0$. So, dim $S = 1$.

In particular, example 3 shows that the topological dimension of R (an infinite line segment) is one. So topological dimension, which is capable of measuring many shapes, agrees with the vector space dimension of R, which measures information.

The following facts are presented without proof:

Theorem: Let dim denote topological dimension.

(i) $\dim R^n = n$.

(ii) If $X_1, X_2, \ldots$ is a countable collection of closed sets (each imbedded in the same R^n for some n) then $\dim \left(\bigcup_{i=1}^{\infty} X_i \right) = \max_i \{\dim X_i\}$

(iii) If $X \neq \phi$ then $\dim (X \times Y) \leq \dim X + \dim Y$.

(iv) If $\dim X = n$, then X can be written as the union of $n + 1$ 0-dimensional subsets of X.

Exercises:

1. What is the topological dimension of the rational numbers, Q?

2. What is the topological dimension of the irrational numbers, R - Q?

3. What is the topological dimension of Q x Q? This is the set of all ordered pairs where both coordinates are rational. What is the dimension of (R -Q) x (R -Q)?

4. A topological space, T, is called disconnected if and only if T can be written as the disjoint union of two nonempty open subsets. The space, T, is totally disconnected if every subset of T consisting of more than one point is disconnected (using the relative topology). The space, T, is connected if it is not disconnected. Is R connected, disconnected, or totally disconnected? How about Q and R - Q?

5. Let S be a finite rectangle in R^2. Cover S with 3 open discs which overlap in the center of S. Is it possible to find 3 sub-discs which cover S but have an empty intersection? What is the dimension of S?

6. What is the dimension of the surface of a cube? Since the surface of a cube and a sphere are topologically equivalent, what is the dimension of a sphere?

I.6 Hausdorff Dimension--A Definition of Fractal

Hausdorff Dimension was the first logically developed tool which could give a non-integer result to be interpreted as the dimension of some shape. With the advent of recent computer graphics, it can be seen that this dimension is useful. The human eye is capable of distinguishing between similar shapes with different Hausdorff dimensions. However, Hausdorff dimension is mathematically difficult to work with. To start, we will discuss some ideas from measure theory.

Lebesque (outer) measure is a way to define the "length" of abstract sets. Essentially, one finds the most efficient open cover of the set, and then finds the length of that open set by adding the lengths of the intervals. We'll restrict ourselves to R for the time being.

Definitions. Let I be an open interval in R. So $I = (a,b)$ for some a and b. The length of I is $b - a$, and is denoted by $|I|$.

Let E be a subset of R. The Lebesque outer measure of E is denoted by $\mathcal{L}(E)$ and $\mathcal{L}(E) = \inf \sum_n |I_n|$ where $\bigcup_n I_n$ is an open set of disjoint component intervals, I_n, and the infimum is taken over all open sets which cover E.

Example 1.

The Lebesque measure of a single point is zero.

Without loss of generality, assume the point is zero as a subset of R. Then, for any $\varepsilon > 0$ we can find an open interval, and hence an open set, which covers zero and has length less than ε. This implies (since ε is arbitrary) that the infimum over all covers is zero, proving the claim.

(A measure, like Lebesque measure, is always defined on a large class of sets which form a "σ-algebra." If one assumes the "Axiom of Choice", then it can be shown that there are subsets of R which are not Lebesque measurable; although every subset is Lebesque outer-measureable.)

For this report, we will assume that the sets we encounter are always measureable. The following are presented without proof:

Theorem.

Let $\mathcal{L}$ denote Lebesque measure. Then:

(i) If $\mathcal{L}(E) = 0$ then every subset of E is measureable and has measure zero.

(ii) If $E_1, E_2, E_3, \ldots$ is any countable collection of disjoint measureable sets, then

$$\mathcal{L}\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \mathcal{L}(E_i).$$

Some immediate consequences of this theorem are that the measure of any interval, (open, closed, or half-open), is computed by subtracting the left endpoint from the right; and that the Lebesque measure of the rational numbers (or any countable set) is zero.

The idea behind Lebesque measure will now be generalized to get Hausdorff (outer) measure. Instead of adding the lengths of all the intervals in the cover, we add a power of the lengths of all the intervals in the cover.

Definitions.

Let $\delta > 0$ and $s > 0$. If E is a subset of R, then the Hausdorff s-dimensional δ-pre-measure of E is:

$$H_\delta^s(E) = \inf \sum_n |I_n|^s$$

where the I_n's are open component intervals of a cover of E and each $|I_n|$ is less than δ. The infimum is taken over all open covers of <u>order δ</u> (that is, each $|I_n| < \delta$).

The <u>Hausdorff s-dimensional</u> (outer) <u>measure of E</u> is

$$H^s(E) = \lim_{\delta \to 0} H^s_\delta(E).$$

Needless to say, this is a very technical definition. However, it is easy to generalize to higher dimensions. (That is, R^m instead of R.) One just defines $|I_n|$ to be the (maximum) diameter of an open set in R^m.

The δ's in the definition are needed for obscure reasons. Basically, $H^s_{\delta'}(E)$ is larger than $H^s_\delta(E)$ if $\delta' < \delta$. Thus, the covers are less efficient as their maximum size decreases.

The s in the definition is the power to which each $|I_n|$ is raised. It will be directly related to the dimension of the set, E. Since s is only restricted to be positive, non-integer dimensions will be possible.

<u>Example 2</u>.

Let E be a line segment of length one. We can assume E = (0,1) in R. The Hausdorff s-dimensional measure of E will be:

a. infinite if $s < 1$,

b. one if $s = 1$, and

c. zero if $s > 1$.

To start, note that if $s = 1$ then Hausdorff 1-dimensional measure is the same as Lebesque measure. So, we need only concern ourselves with parts a and c.

Part a. Let $s < 1$.

Choose a large $M > 0$. We'll show that there is a small δ so that $H^s_\delta(E) > M$, and that as M gets larger, δ gets smaller. This will imply that $\lim_{\delta \to 0} H^s_\delta(E) = \infty$, completing the proof of this part.

So, we choose a large integer, k, such that $k^{1-s} > M$ and then let $\delta = \frac{1}{k}$. (Since $s < 1$, as M increases so will k, and hence δ will get smaller--approaching zero.) The most efficient cover of order δ (of E) will be less efficient than k open intervals, each of length $\delta = \frac{1}{k}$. So,

$$H^s_\delta(E) \geq \sum_{n=1}^{k} \left(\frac{1}{k}\right)^s > \sum_{n=1}^{k} \left(\frac{M}{k}\right) = M.$$

This shows that the Hausdorff s-dimensional measure of E is infinite when $s < 1$.

Part c. Let $s > 1$.

Choose a small number $\varepsilon > 0$. We'll show that there is a small δ so that $H^s_\delta(E) < \varepsilon$, and that as ε gets smaller, so does δ. This will imply that $\lim_{\delta \to 0} H^s_\delta(E) = 0$.

Let $\varepsilon > 0$, and choose a large integer k so that $k^{1-s} < \varepsilon$, and $\delta = \frac{1}{k}$. (Since $s > 1$, as ε decreases, δ will decrease as well.) The most efficient cover of order δ will be approximately as efficient as k intervals of length $\frac{1}{k}$. Thus,

$$H^s_\delta(E) \approx \sum_{n=1}^{k} \left(\frac{1}{k}\right)^s < \sum_{n=1}^{k} \left(\frac{\varepsilon}{k}\right) = \varepsilon.$$

This shows that the Hausdorff s-dimensional measure of E is zero when $s > 1$, completing the proof.

Example 2 displays the type of behavior that Hausdorff s-dimensional measure always exhibits. For subsets of R, when $s > 1$ the measure will be zero, and when $s = 1$ the measure will agree with Lebesque measure. If the Lebesque measure is positive, then for smaller values of s, the Hausdorff s-dimensonal measure will always be infinite. (This can happen even if the Lebseque measure is zero--see section I.8.)

These qualities are used to define Hausdorff dimension:

Definitions.

Let E be a subset of R. The Hausdorff dimension of E is the infimum of the set of s such that the Hausdorff s-dimensional measure of E is zero.

This definition is extended to R^m by using the diameter of I_n instead of the length of I_n. So, $|I_n| = \sup \{d(x,y): x,y \in I_n\}$ where "d" denotes distance in R^m, and I_n is now any open set. Then the Hausdorff dimension of subsets of R^m is the same as that above.

Example 2 shows that the Hausdorff dimension of the open interval (0,1) is one. This agrees with both intuition and topological dimension. In fact, the Hausdorff dimension of a point (and any countable set) is zero (also in agreement), and the Hausdorff dimension of R^n is n. This leads us to a definition of "fractal."

Definition. A subset, E, of R^n is a fractal if the topological dimension of E is different than the Hausdorff dimension of E.

The first fractal we will encounter will be the "Cantor Set" in section I.8. But, since Hausdorff dimension is hard to work with, we'll define a "similarity dimension" in the next section. This similarity dimension will be easy to calculate on special sets, and will agree with the Hausdorff dimension on such sets.

Exercises

1. Show that the Hausdorff dimension of a single point is zero.
2. What is the Lebesque measure of the irrational numbers between zero and one?
3. What is the Hausdorff dimension of the irrational numbers between zero and one?
4. Is the set of irrational numbers between zero and one a fractal?
5. What is the Hausdorff dimension of $E = \{(x,y): 0 \leq x,y \leq 1\}$?

I.7 Similarity Dimension

Similarity dimension is an easier tool for measuring the (Hausdorff) dimension of special sets. These sets are "self-similar". We will use the following definition for now:

Definitions.

For the following recursive procedure, N is a positive integer and r is a real number, $0 < r < 1$.

(i) Start with a generating set, E_0, with one component.

(ii) Given E_n, with N^n components, form E_{n+1} by replacing each component of E_n by N new components so that the diameter of each new component is r times the diameter of the old component. E_{n+1} will have N^{n+1} components.

If this recursive process has a limit set, E, which uniquely depends upon the choices for N and r, then E will be called a self-similar set.

The term, "component", used above, is intentionally vague. But, whatever is meant by component in part (i) is to be used in part (ii) as well. "Component" could be replaced by "shape". Note that there is no restriction on the R^m in which this process takes place.

For example, E_0, could be a line segment. Then, each line segment would be replaced by N new line segments, r times the size. Or, one could use triangles, spheres, pyramids, helices, etc.

We'll now motivate the formula used in similarity dimension. It is equivalent to Hausdorff dimension for self-similar sets...

In the recursive process, E_1 could be covered by N open sets, each of diameter r. (We assume that the diameter of E_0 is 1.) Then, if $\delta > r$, $H^s_\delta(E_1) = N(r)^s$. Note that for E_2, the diameters are r^2, but there are N^2 components, so $H^s_\delta(E_2) = (Nr^s)^2$, and in general, $H^s_\delta(E_n) = (Nr^s)^n$. Now, as δ

goes to zero, we have n going to infinity, (so that r^n is less than δ), but this will not give a nonzero or noninfinite limit unless $Nr^s = 1$. However, we're looking for this value of s, say s_0. When s is bigger than s_0, the Hausdorff s-dimensional measure will be zero, and when s is less than s_0, the Hausdorff s-dimensional measure is infinite. So, s_0 is the Hausdorff dimension of E and satisfies $Nr^{s_0} = 1$. (Or, $s_0 = \frac{\log N}{\log(1/r)}$. Note: E is imbedded in R^m, for some m, so s_0 can never be greater than m.)

Definition. The similarity dimension of a self-similar set, E, is $\frac{\log N}{\log(1/r)}$ where N and r are as defined in the recursive process. This is also the Hausdorff dimension. The formula fails if it generates a dimension larger than that of the space in which E is imbedded.

Exercises.

1. Define a point as follows: Start with a line segment, and replace any line segment by a smaller (subset) line segment one-half the size. Then $N = 1$ and $r = \frac{1}{2}$. What is the dimension of a single point. Is a point uniquely defined in terms of N and r?

2. Restricting N as a positive integer and r as a real number, $0 < r < 1$, what is the range of the function: $s(N,r) = \frac{\log N}{\log(1/r)}$?

3. Suppose you observe a coastline from orbit. You digitize a picture of it and count a total of 100 pixels on the coastline between two fixed points. The 100 pixels given an approximate length of 100 kms. You are now in an airplane over the same coastline and notice a new digitized image has twice the resolution. After counting pixels on the coastline you now get an approximate coastline length (between the same points) of 121 kms. Finally, with a good map with twice the previous resolution you approximate the length to be 146 kms. What approximate dimension would model this coastline?

I.8 The Cantor Set

This will be the first section which is dedicated to applying our tools to examples of fractals. The Cantor Set is one of the most famous "weird" sets in all of mathematics. It was originally developed by Georg Cantor (1845-1918) to show properties of cardinality: that it is possible for a "small" set to be "uncountable". We will give three equivalent formulations of this set.

Formulation 1. (Self-similarity.) Let E_0 be the closed interval, [0,1], in R. Choose N = 2 and $r = \frac{1}{3}$. Given a line segment (component) in E_n, remove the middle (open) $\frac{1}{3}$ and replace the segment with the 2 closed end-segments of length $\frac{1}{3}$ times the original:

Formulation 2. (Closed Set.) From the interval [0,1] in R, remove the open interval which is the middle one-third. For each closed interval remaining repeat this process ad infinitum.

Formulation 3. (Tertiary numbers.) All real numbers may be written in a base 3 expansion using the digits 0, 1, and 2. This expansion is unique, up to repeating 2's which are equivalent to a 1 followed by repeating zeros. (This is analogous to $.\overline{99} = 1.\overline{0}$.) The Cantor Set is the collection of all real numbers between zero and one (inclusive) which can be written in base 3 using only the digits 0 and 2. (So .1 (base 3) is in the Cantor Set since .1 (base 3) = $.0\overline{2}$ (base 3).)

It isn't too hard to see that formulations 1 and 2 give the same set. But it is rather remarkable that formulation 3 is also the same set!

From formulation 1, we can calculate the Hausdorff dimension (similarlity dimension) of the Cantor Set to be log(2)/log(3). This is a noninteger number, so it cannot agree with the topological dimension. This shows that the Cantor Set is a fractal.

Definition. A subset, E, of a topological space, T, is nowhere dense if and only if for every open set, U, in T, there is an open subset of U which does not intersect E_0.

Theorem. Let C denote the Cantor Set.

(i) The topological dimension of C is zero.

(ii) The Lebesque measure of C is zero.

(iii) C is an uncountable set.

(iv) C is a closed set, with no isolated points.

(v) C is nowhere dense.

Proofs.

(i) C is a totally disconnected set by construction. Let $G_1, \ldots, G_p$ be open sets so that $\bigcup_{i=1}^{p} G_i$ covers C. Then we can choose open subsets $H_i \subseteq G_i$ which cover C and are at most abutting. That is, the H_i's do not overlap. Thus, the topological dimension of C is zero.

(ii) The complement of C inside [0,1] is an open set. We'll denote it by P. We'll find the Lebesque measure of P by adding the lengths of all its intervals:

$$\mathcal{L}(P) = \sum_{i=1}^{\infty} 2^{i-1}\left(\frac{1}{3}\right)^i,$$

since at the ith level of construction, one removes 2^{i-1} intervals, each of length $\left(\frac{1}{3}\right)^i$. Thus,

$$\mathcal{L}(P) = \sum_{i=1}^{\infty} \frac{1}{2}\left(\frac{2}{3}\right)^i = \frac{1}{2}\ \frac{2/3}{1-2/3} = 1.$$

Since P_1 has measure one inside [0,1] and $C = P^c$, $\mathcal{L}(C) = 0$.

(iii) We'll prove this part by using formulation 3. The tertiary numbers between 0 and 1 which only contain the digits 0 and 2 are in 1 - 1 correspondence with the binary numbers between 0 and 1. Since there are uncountably many (binary) numbers between 0 and 1, C is also uncountable.

(iv) C is a closed set since its complement, P, is open. Also, the endpoints of the intervals which were removed from C are elements of C. These endpoints form a countable subset of C, but they are dense in C; i.e., each point of C is a limit point of the endpoints.

(v) Given any open set, U, in [0,1], there is an open interval in P which is completely contained in U. Since $P \cap C = \phi$, C must be nowhere dense.

Thus, C is a perfect (closed with no isolated points), null (zero Lebesque measure), nowhere dense fractal with Hausdorff dimension equal to log(2)/log(3).

It is easy to construct Cantor-type sets by removing the middle p, $(0 < p < 1)$, instead of the middle one-third, of each remaining interval. The resulting set will share most properties of the Cantor set, except its Hausdorff dimension and self-similarity.

In fact, it is possible to construct a Cantor-type set which has zero Hausdorff dimension. This means its topological and Hausdorff dimension agree, so it is not a fractal. This points out a serious flaw in the definition of "fractal." A more general definition will be given later.

Exercises.

1. Form a Cantor-type set by removing the middle p, $(0 < p < 1)$, of each remaining interval instead of the middle one-third. What is its Hausdorff dimension? For $0 < p < 1$, what is the range of values for this dimension.

2. Given any open set, U, in R, what is the Hausdorff dimension of U? What is the Hausdorff dimension of the complement of any Cantor-type set?

3. If you divide an interval into 4 pieces and remove the second one, then repeat this process ad infinitum, you obtain another self-similar closed set. In general, let $N = n - 1$ and $r = \frac{1}{n}$. For such self-similar sets, state an analogous theorem to that for the Cantor set. (Include this set's Hausdorff dimension.) If you replace $N = n - 1$ by $N = n - m$, what values of m make sense? What is the range of achievable Hausdorff dimensions?

4. In R^2, divide a square into 9 subsquares and remove the middle one. Will this recursive process generate a fractal? What is its Hausdorff dimension. (Remember to use diameter lengths when calculating r.)

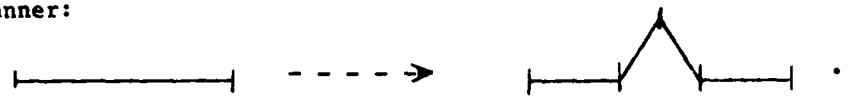
5. In R^3, divide a cube into 27 subcubes. At each step, (level of recursive process), remove 7 of these subcubes. Is the Hausdorff dimension effected by which 7 are removed? Could one remove a different (or random) 7 at each level?

I.9 Koch Snowflakes

Some of the first "monster curves" or "snowflakes" were constructed by Helge von Koch. We will look at a few examples of self-similar shapes in these categories:

Example i.

Let the generating set, E_0, be a line segment. Use $N = 4$ and $r = \frac{1}{3}$ in the following manner:



Whenever you encounter a line segment, replace it by four line segments in the above formation. The dimension of the resulting "curve" is log 4/log 3 which is larger than 1.

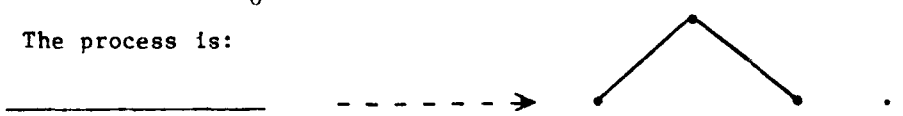
At the n^{th} stage, the length of E_n is $4^n(\frac{1}{3})^n = (\frac{4}{3})^n$. This quantity approaches infinity as n approaches infinity, so the curve must have infinite length. On the other hand, the entire curve will fit inside a bounded square. This is not a contradiction; for example, if we tried to compute the 1-dimensional length of a 2-dimensional square, we would get an infinite answer.

The Hausdorff s-dimensional measure is exactly the tool necessary to compute "lengths" (or measures) of s-dimensional shapes. It is necessary to adapt the measurement to the dimension of the shape you are measuring. The measurement could still be infinite, or zero, but it would have more meaning: just as it means more to say that a line segment is 2 units long rather than it has zero area.

Unfortunately, it is difficult to calculate the Hausdorff s-dimensional measure of a shape (where s is the shape's dimension). This measurement is also dependent on the type of sets one allows in the cover. For example, in R^2, if we only allow open discs, we will get a different measurement than when wee allow any open sets to be used. For this reason, we note that such measurements exist and have meaning, but won't be used in this report.

Example 2.

The generating set, E_0, will be the same as in Example 1, but N = 2 and $r = \frac{1}{\sqrt{3}}$. The process is:



The self-similar shape resulting from this recursive process has dimension $\log 2/\log \sqrt{3} = \log 4/\log 3$, so the dimensions of Examples 1 and 2 are the same (in fact, they are more directly related than that--see the exercises).

Since in both of the examples above each shape is connected (not disconnected) and lies in R^2, their topological dimensions are at least one. It is a bit harder to show that their topological dimensions are actually equal to one. But for any shape, E, the topological dimension of E is always less than or equal to the Hausdorff dimension of E. Regardless, the shapes in examples 1 and 2 are fractal curves.

Example 3.

The generator, E_0, will still be a line segment. Now, N = 3 and $r = \frac{1}{2}$. The process is:

The height, h, is $\frac{\sqrt{3}}{4}$ times the length of the segment on the left. Thus, if E_0 has length 1, the farthest displaced point from E_0 in the limit set, E, will be $\sum_{n=1}^{\infty} \left(\frac{\sqrt{3}}{4}\right)^n = \frac{3 + 4\sqrt{3}}{13}$ units away. The Hausdorff dimension of this fractal curve is log 3/log 2.

Exercises.

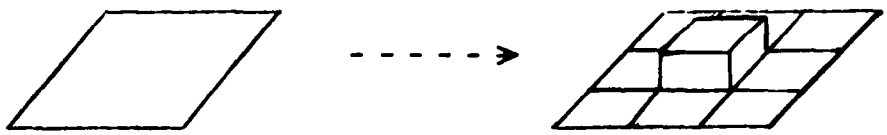
1. In example 2, if $r = \frac{1}{\sqrt{2}}$, the Hausdorff dimension of the resulting set is 2. This is an area filling curve. If this process is done in R^2 and $\frac{1}{\sqrt{2}} < r < 1$, does the predicted Hausdorff dimension make sense?

2. As in example 2, let $N = 2$ and $r = \frac{1}{\sqrt{3}}$. Each time the process is done, flip a coin to determine whether the triangle will displace up or down. Does the dimension change from that of example 2? Give a rough sketch of a resulting E.

3. For $N = 3$ and a fixed r, $\frac{1}{3} < r < \frac{1}{\sqrt{3}}$, there are more degrees of freedom in a random process than there were in exercise 2. Explain.

4. If we take example 1 and apply it to the 3 sides of an equilateral triangle, we get a "snowflake". What is its dimension? Does this change if we apply it to the sides of a square?

5. Let E_0 be a square imbedded in R^3. Define $N = 13$ and $r = \frac{1}{3}$ as follows:



What is the Hausdorff dimension of E?

6. Exercise 5 may be applied to the 6 sides of a cube. What shape does it resemble?

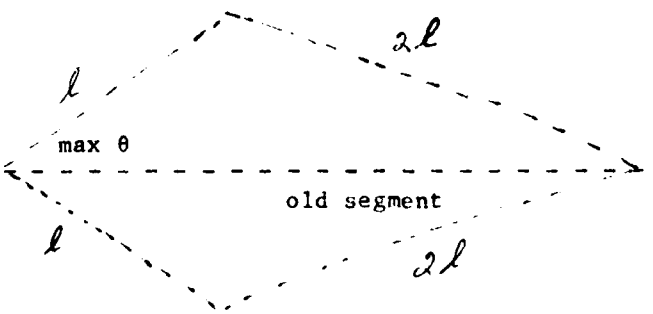
7. Show that a countably infinite set of points on the curves generated in examples 1 and 2 are the same.

I.10 Random Coastlines

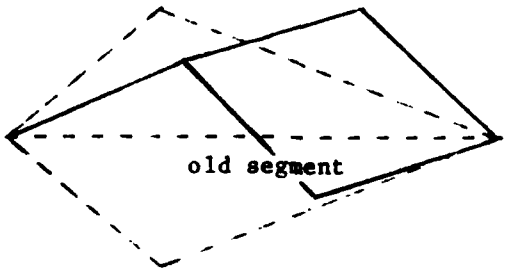
Exercise 3 in section I.9 contains an idea we can use to generate random fractal curves of any specified dimension between one and two. Since self-similarity does not require that the same scheme be used at each level of the recursive process, (only N and r need be fixed), we can randomly change the scheme to obtain a random fractal curve.

If $N = 3$ and $\frac{1}{3} \leq r \leq \frac{1}{\sqrt{3}}$, by specifying r, it is possible to achieve any dimension between one and two. In example 3 of section I.9, $r = \frac{1}{2}$ so the resulting dimension was log 3/log 2. If you want the dimension to be d, then choose $r = N^{-1/d}$ where $N = 3$.

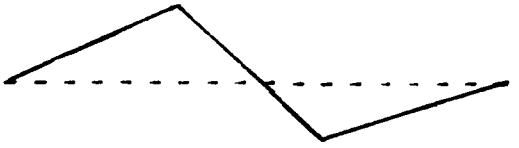
Since we want our curve to be connected, at each stage we join 3 segments where there was only one. These segments have fixed length, ℓ:



In the above diagram, we can place the first segment anywhere within the maximum arc about the left endpoint. (See the exercises.)



Once this segment has been placed, there are exactly 2 choices for the placement of the second segment (unless the first was at $\pm$ max θ, in which case there is only one choice). The placement of the third segment is determined.



So, this algorithm requires a random number between -max θ and max θ for placement of the first segment; and a coin toss (0,1 random variable) for placement of the second segment which also determines placement of the third segment. The continuous random variable in the process allows for greater variability than a single coin toss at each step (which is what happens if N = 2).

The four pages of figures which follow make it clear that this process can be used to model coastlines. At first glance, this fact seems to be of little use. However, the fact is useful from a "first-principles" point of view, leading us to believe that coastlines are self-similar objects. Thus, any valid geological theory on the evolution of coast-lines would need to address some similarity at different scales.

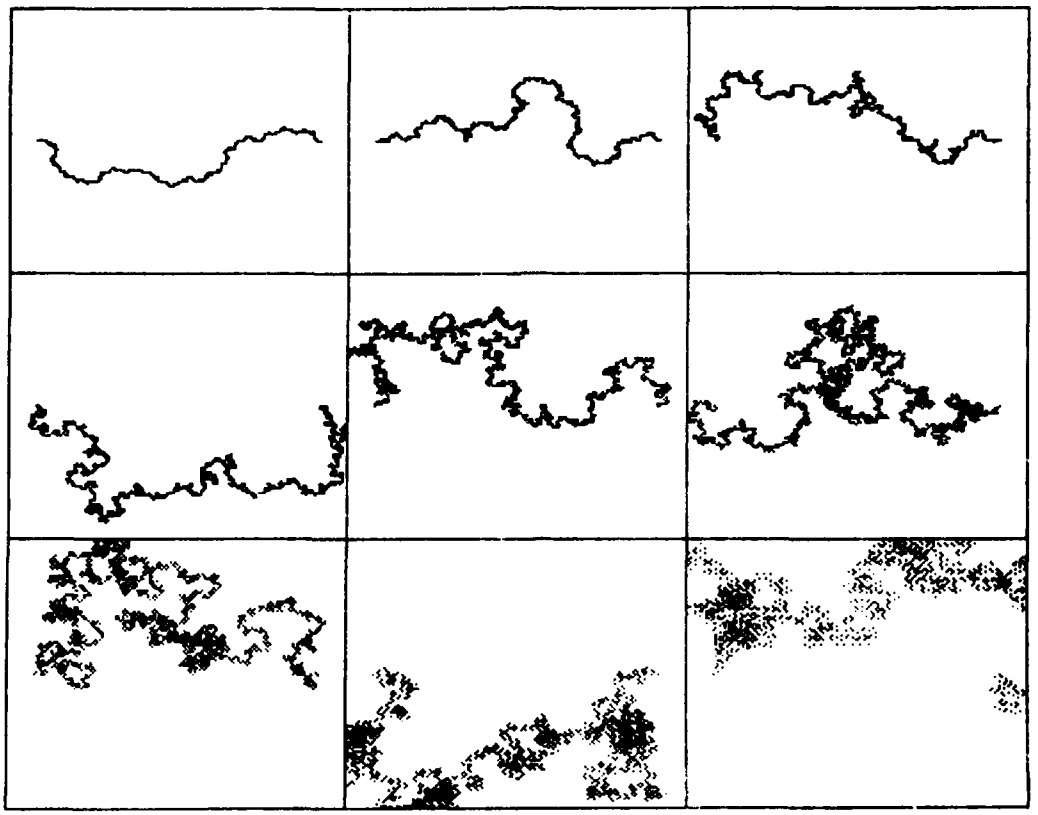


Figure 2

"The nine frames in this figure depict random coastlines generated with an algorithm based on exercise 2 in section I.9. The ratio, r, is varied so that dimensions of 1.1, 1.2, ..., 1.9 are realized. For each picture, the probability of displacing up is 1/2. 2049 points are graphed per frame. We can see the higher dimension fractals are area filling.

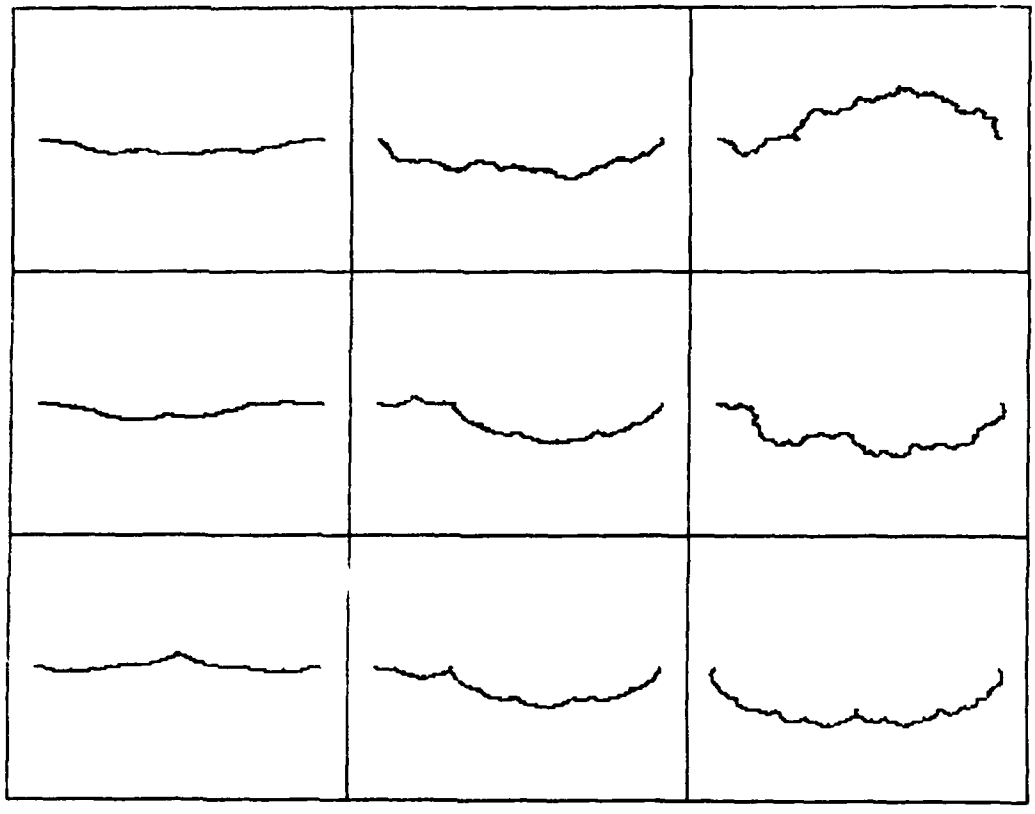


Figure 3

"In each row of this figure, dimensions are 1.01, 1.05, and 1.1 (from left to right). The same algorithm was used as in figure 2, but now the probabilities of curving down are .5, .7, and .9 for the top through bottom row, respectively. 1025 points were plotted in each figure."

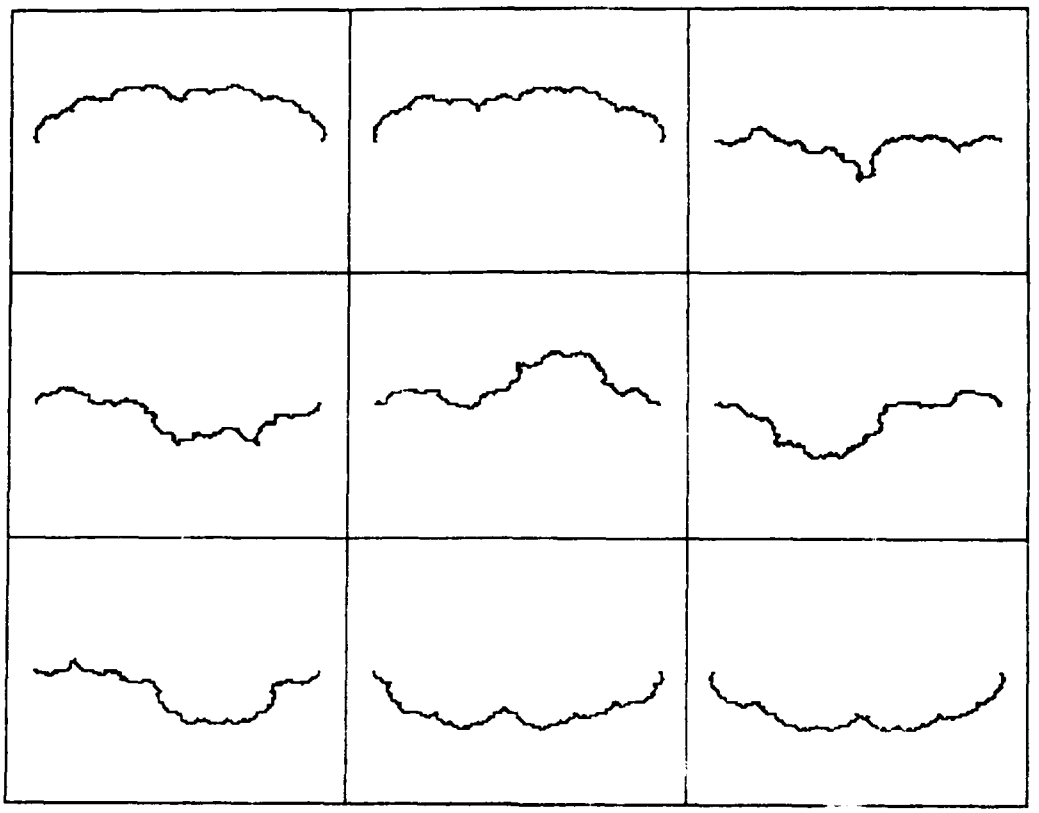


Figure 4

"Each random fractal in this figure has dimension 1.1. The probabilities of curving down vary from .1 to .9."

"On the following page: All 18 frames in figure 5 use the 3 segment algorithm developed in this section. In the first 3 rows, a uniform distribution is used for the angle of the first segment and .5 is used for the next two segments (just like a 2-segment algorithm). The last 3 rows all depict a dimension of 1.1. The last 2 rows use a truncated normal distribution with small and large standard deviation (respectively) and probabilities of .1 and .9, respectively for the last 2 segments."

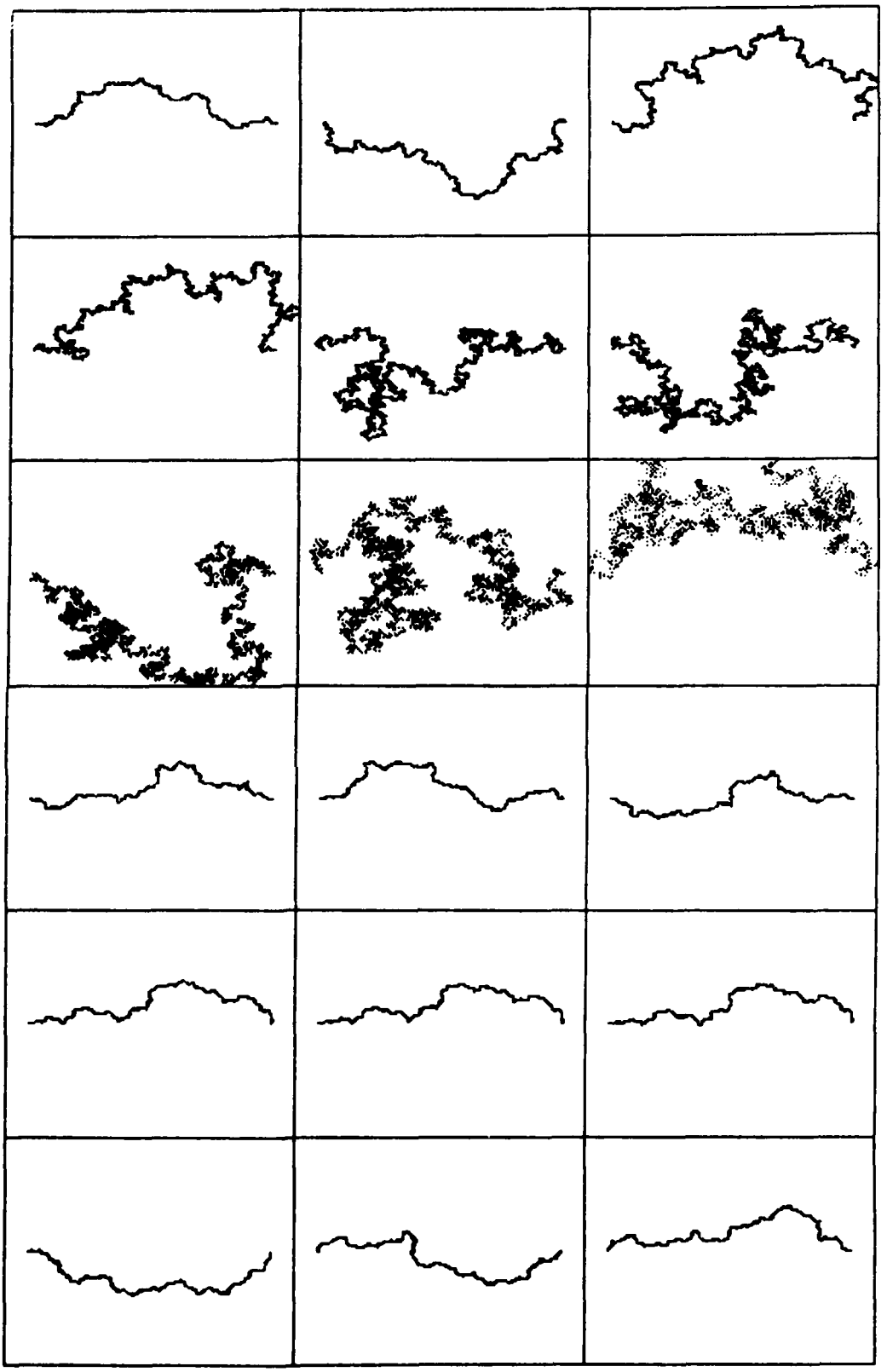


Figure 5

Unfortunately, the algorithm described in this section is not too useful outside of graphic simulation. A much more applicable idea would be to use self-similarity in order to decrease the amount of data required to depict an actual geographic object. That is, given a mountain, is it possible to generate an accurate image of this mountain without using the millions of bits of information contained in a high resolution photograph? If it were, then the image could be transmitted to another person (or machine) in a fraction of the usual time.

Exercises.

1. Given N = 3 and fixed r, with $\frac{1}{3} < r < \frac{1}{\sqrt{3}}$, what is the value of max θ as described earlier in this section?

2. If the first segment makes an angle of max θ with the left endpoint of the old segment, what angle will the third segment make with the right endpoint? Note that this value was used as the standard deviation of the (truncated) normal distribution in some of the figures in this section.

3. Construct an algorithm similar to the one in this section, but use N = 4.

I.11 Another Approach - IFS

The definition of a self-similar set given in I.7 is neither exact nor formal. It is useful from an intuitive point of view, but not too useful mathematically. In this section we will develop the framework necessary to study Iterated Function Systems (IFS).

The paper "Fractals and Self-Similarity" by John Hutchinson gives a complete discussion of the ideas behind IFS. Michael Barnsley coined the term IFS and has exploited the mathematics for applications for DARPA, AFOSR, and his own company, Iterated Systems, Inc.

Three topics will be introduced in this section: "compact sets", "the Hausdorff metric", and "contraction maps". It will help to use the following relations:

a. Compact sets will be pictures, images, or real shapes.

b. The Hausdorff metric will measure how close two compact sets (pictures) are to being the same.

c. Contraction maps will act as lenses which focus any starting image into a determined compact set (picture)--sort of like a kaleidoscope.

In practice, the contraction maps will be repeatedly applied to any generating set (image). Each application will bring a new image which is closer to the limiting image (compact set). This process can be used to generate fractals like those we have seen earlier, or extremely detailed pictures of nature, including human faces. (The less self-similar seeming the picture, the greater number of contraction maps are necessary to generate it.)

We will start with the idea of a compact set.

Definition. Let (T, Ω) be a topological space. A subset of T, F, is compact if and only if for all collections $\{A_\alpha : \alpha \in I, A_\alpha \in \Omega\}$ such that $\bigcup_{\alpha \in I} A_\alpha \supseteq F$, there is a finite subcollection $\{A_{\alpha_i} : i = 1, \ldots, n, \alpha_i \in I\}$ such that $\bigcup_{i=1}^{n} A_{\alpha_i} \supseteq F$ also.

In words, a set is compact if and only if given an arbitrary open cover, there is a finite subcover.

Example 1. A point is compact.

Let $0 \in R$. If there is an arbitrary collection of open sets in R which cover $\{0\}$, then at least one of those sets must actually contain 0. This set is a finite subcover, so $\{0\}$ is compact.

Example 2. An open interval is not compact.

Let $(0,1)$ be a subset of R. We will construct an open cover of $(0,1)$ which has no finite subcover:

Define $G_n = (\frac{1}{n}, 1)$. Then $\bigcup_{n=1}^{\infty} G_n = (0,1)$ so $\{G_n\}$ is an open cover of $(0,1)$. If there were a finite subcover from $\{G_n\}$, then there would be a maximum index, say N, in that subcover. But $G_N = (\frac{1}{N}, 1)$, and since no index in the subcover is larger than N, the points between 0 and $\frac{1}{N}$ are not covered by this finite collection. Thus, there is no finite subcover and $(0,1)$ is not compact.

Definition. A subset, B, of R^n is bounded if and only if $\sup \{d(x,0): x \in B\}$ is finite where d is Euclidean distance. That is, B is bounded if and only if B will fit inside some n-ball of finite radius.

We will state the following useful theorem without proof:

Theorem. A subset of R^n is compact if and only if it is closed and bounded.

Example 3. The Cantor Set is compact since it is closed and bounded in R.

Example 4. All the fractal curves generated in section I.10 are compact, since they are also closed and bounded.

Example 5. Any image on a monitor is compact since it consists of a finite number of points (pixels).

Example 6. Any observable shape of matter is compact. (One would need to be careful with particle/wave duality at a very small scale.)

The next two topics require a review of metric spaces.

Definition. If S is a set and $d: S \times S \to R$, then (S,d) is a metric space if and only if the following conditions are satisfied:

(i) for all $x,y \in S$, $d(x,y) \geq 0$ and $d(x,y) = 0$ if and only if $x = y$.
(ii) for all $x,y \in S$, $d(x,y) = d(y,x)$.
(iii) for all $x,y,z \in S$, $d(x,z) \leq d(x,y) + d(y,z)$.

Remarks. A metric space abstracts the idea of distance, embodied in the function, d. The first two requirements are straightforward; the third requirement says that it's always shorter to go from one point directly to another. Also, every metric space generates a natural topology in the following manner:

A point, x, is an interior point of A (a subset of the metric space), if and only if there is an $\varepsilon > 0$ so that if $d(x,y) < \varepsilon$ then $y \in A$. With this definition, a set is open if and only if all of its elements is an interior point.

This development shows that a topological space is more abstract than a metric space since every metric space is a topological space, but not vice versa.

Example 7. (R^n,d) is a metric space where $d((x_1, \ldots, x_n),(y_1, \ldots, y_n)) = \left[\sum_{i=1}^{n} (x_i - y_i)^2 \right]^{1/2}$.

This Euclidean metric, d, generates the usual topology on R^n. (Note n = 1, 2, 3...)

We are now in a position to define the Hausdorff metric.

Definition. Let n be a fixed natural number, and $S = \{F \subseteq R^n: F \text{ is compact and nonempty}\}$.

If x is any point in R^n and $A \subseteq R^n$, extend the usual Euclidean metric to $d(x,A) = \inf\{d(x,y): y \in A\}$. So, d(x,A) measures the closest distance between the point and the set. Now extend d to A and B, (subsets of R^n), as d(A,B) is the maximum of $\sup\{d(x,B): x \in A\}$ and $\sup\{d(y,A): y \in B\}$.

d is the Hausdorff metric.

We need to show that (S,d) is a metric space.

Theorem. (S,d), described above, is a metric space.

Proof. Fix $n \in N$ and let A, B, and C be compact sets in R^n, and thus elements of S.

(i) $d(A,B) \geq 0$ since the Euclidean metric is nonnegative. d is finite valued since both A and B are bounded sets. Assume A = B. Then, for all $x \in A$, d(x,B) = 0. Thus, $\sup\{d(x,B): x \in A\} = 0$. Similary, $\sup\{d(y,A): y \in B\} = 0$. Therefore, d(A,B) = 0. Now assume d(A,B) = 0. Then, $\sup\{d(x,B): x \in A\} = 0$, implying that each d(x,B) = 0, since they're all nonnegative. Thus, $\inf\{d(x,y): y \in B\} = 0$ for each x. This implies that x is a limit point of B. Since B is closed, $x \in B$ for each $x \in A$, implying $A \subseteq B$. Similarly, $B \subseteq A$, which shows that A = B and completes part (i) of the definition of a metric space.

(ii) By definition,

$$d(A,B) = \max \{\sup_{x\in A} d(x,B), \sup_{y\in B} d(y,A)\}$$

$$= \max \{\sup_{y\in B} d(y,A), \sup_{x\in A} d(x,B)\}$$

$$= d(B,A)$$

(iii) $\sup_{z\in C} d(z,A) = \sup_{z\in C} \inf_{x\in A} d(x,z)$.

Since B is a closed set, there is a y_z for which $d(y_z,z) = \inf_{y\in B} d(y,z)$

$= d(z,B)$. Thus, $\sup_{z\in C} d(z,A) \leq \sup_{z\in C} \inf_{x\in A} [d(x,y_z) + d(y_z,z)]$

$$= \sup_{z\in C} \inf_{x\in A} [d(x,y_z) + d(z,B)]$$

$$= \sup_{z\in C} [\inf_{x\in A} d(x,y_z) + d(z,B)]$$

$$= \sup_{z\in C} [d(y_z,A) + d(z,B)]$$

$$= d(y_z,A) + \sup_{z\in C} d(z,B)$$

$$\leq \sup_{y\in B} d(y,A) + \sup_{z\in C} d(z,B).$$

By an exactly similar argument,

$$\sup_{x\in A} d(x,C) \leq \sup_{y\in B} d(y,C) + \sup_{x\in A} d(x,B).$$

Therefore,

$$d(A,C) = \max \{\sup_{x\in A} d(x,C), \sup_{z\in C} d(z,A)\}$$

$$\leq \max \{\sup_{y\in B} d(y,C) + \sup_{x\in A} d(x,B), \sup_{y\in B} d(y,A) + \sup_{z\in C} d(z,B)\}$$

$$\leq \max \{\sup_{y\in B} d(y,C), \sup_{z\in C} d(z,B)\} + \max \{\sup_{x\in A} d(x,B), \sup_{y\in B} d(y,A)\}$$

$= d(B,C) + d(A,B)$, completing part (iii) and showing that (S,d) is a metric space.

Example 8. In R, the (Hausdorff) distance between the Cantor Set and the interval [0,1] is $\frac{1}{6}$.

Let A be the Cantor Set and B = [0,1]. Since $A \subseteq B$, d(x,B) = 0 for each $x \in A$. Thus, $\sup_{x \in A} d(x,B) = 0$. If $y \in B$ then d(y,A) > 0 when $y \notin A$. The largest part of B which does not contain points of A is the interval $(\frac{1}{3}, \frac{2}{3})$. It is easy to see that $y = \frac{1}{2}$ is the point in B which is farthest from any point in A, so $d(\frac{1}{2}, A) = \frac{1}{6}$. Thus, $d(A,B) = \frac{1}{6}$.

Example 9. In R^2, the (Hausdorff) distance between $A = \{(x,y): x^2 + y^2 \leq 1\}$ and $B = \{(x,y): y = 0 \text{ and } -2 \leq x \leq 2\}$ is 1.

It is easy to check that $\sup_B d((x,y),A) = 1$. Also, $\sup_A d((x,y),B) = 1$. Thus, d(A,B) = 1.

We will now define contraction mappings.

Definition. If $f: S \to S$ where (S,d) is a metric space, then f is a contraction mapping if and only if there is a real constant, r, so that $0 \leq r < 1$, and for all $x,y \in S$, $d(f(x), f(y)) \leq rd(x,y)$.

Intuitively, a contraction mapping decreases the distance between any two points. Consequently, every contraction mapping is continuous, and every contraction mapping has a unique "fixed point." That is, if f is a contraction mapping on S, then there is a unique point, x_0, so that $f(x_0) = x_0$. The proofs of these facts are left for the exercises.

Example 10. If $f: R \to R$ by $f(x) = \frac{x}{3}$, then f is a contraction map.

Let $r = \frac{1}{3}$. Then, for any $x,y \in R$, $d(f(x), f(y)) = |f(x) - f(y)| = |\frac{x}{3} - \frac{y}{3}| = \frac{1}{3}|x - y| = rd(x,y)$. Zero is the fixed point.

Example 11. Define $f: R^2 \to R^2$ by $f(x,y) = \left(\frac{1}{4}[\sqrt{3}x - y], \frac{1}{4}[\sqrt{3}y + x]\right)$. Then f is a contraction mapping with $r = \frac{1}{2}$, and fixed point, (0,0).

Let (x_1,y_1) and (x_2,y_2) be points in R^2.

$$d(f(x_1,y_1), f(x_2,y_2)) =$$

$$= d\left(\left(\tfrac{1}{4}[\sqrt{3}\,x_1 - y_1], \tfrac{1}{4}[\sqrt{3}\,y_1 + x_1]\right), \left(\tfrac{1}{4}[\sqrt{3}\,x_2 - y_2], \tfrac{1}{4}[\sqrt{3}\,y_2 + x_2]\right)\right)$$

$$= \sqrt{\left(\tfrac{1}{4}[\sqrt{3}\,(x_1 - x_2) - (y_1 - y_2)]\right)^2 + \left(\tfrac{1}{4}[\sqrt{3}\,(y_1 - y_2) + (x_1 - x_2)]\right)^2}$$

$$= \tfrac{1}{4}\sqrt{3(x_1 - x_2)^2 + (y_1 - y_2)^2 + 3(y_1 - y_2)^2 + (x_1 - x_2)^2}$$

$$= \tfrac{1}{2}\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

$$= \tfrac{1}{2}\, d((x_1, y_1), (x_2, y_2)).$$

It is easy to see that zero is the fixed point. This function rotates a point by $\frac{\pi}{6}$ radians, counterclockwise about the origin, and decreases its distance from the origin by $\frac{1}{2}$. Thus, repeated applications of f will take a point through a spiral trajectory into the origin.

Definitions. Let $f_1, \ldots, f_m$ be contraction maps on R^n. Then $\{f_i: i = 1, \ldots, m\}$ is called an iterated function system (IFS).

If $A \subseteq R^n$, then $f_i(A) = \{y \in R^n$: there is an $x \in R^n$ and $f_i(x) = y\}$ for each $i = 1, \ldots, m$; $f_i^0(A) = A$ and $f_i^{k+1}(A) = f_i(f_i^k(A))$.

We will let f denote the collection $\{f_i: i = 1, \ldots, m\}$ as follows:

$$f(A) = \bigcup_{i=1}^{m} f_i(A),$$

and $f^k(A)$ is the k^{th} iterate of f.

The following fact is presented without proof:

Theorem. If $f = \{f_i : i = 1, \ldots, m\}$ is an interated function system of contraction maps on R^n, then there is a unique compact set, $F \subseteq R^n$, such that $f(F) = F$ and for any nonempty set, $A \subseteq R^n$, $\lim_{n\to\infty} f^n(A) = F$ in the Hausdorff metric.

Example 12. Let $f = \{f_1, f_2\}$ where each $f_i: R \to R$ by $f_1(x) = \frac{x}{3}$, $f_2(x) = (x+2)/3$. Then, the fixed point of f is the Cantor Set.

First, we'll show that $\lim_{n\to\infty} f^n([0,1]) = C$, where C is the Cantor Set:

$$f_1([0,1]) = \{y: \text{there is an } x \in [0,1] \text{ and } f_1(x) = y\} = [0,\tfrac{1}{3}].$$

Also, $f_2([0,1]) = [\frac{2}{3},1]$.

Thus, $f([0,1]) = f_1([0,1]) \cup f_2([0,1]) = [0,\frac{1}{3}] \cup [\frac{2}{3},1]$. We can see that this is the first level of the construction of C in section I.8.

Now, we'll calculate $f^2([0,1]) = f(f[0,1])$:

$$f_1([0,\tfrac{1}{3}] \cup [\tfrac{2}{3},1]) = [0, \tfrac{1}{9}] \cup [\tfrac{2}{9}, \tfrac{1}{3}].$$

$$f_2([0,\tfrac{1}{3}] \cup [\tfrac{2}{3},1]) = [\tfrac{2}{3}, \tfrac{7}{9}] \cup [\tfrac{8}{9},1].$$

So, $f^2([0,1]) = [0,\frac{1}{9}] \cup [\frac{2}{9}, \frac{1}{3}] \cup [\frac{2}{3}, \frac{7}{9}] \cup [\frac{8}{9},1]$. In general, f_1 will take an interval and move both the endpoints $\frac{1}{3}$ closer to the origin. f_2 will do the same thing, and then add $\frac{2}{3}$ to every point in that interval, shifting it to the right by $\frac{2}{3}$. One should be able to see that this is equivalent (when repeated ad infinitum) to both formulations 1 and 2 in section 1.8. Thus, $\lim_{n\to\infty} f^n([0,1]) = C$.

Let's try to show that C is in fact a fixed point of f. To do this, we'll use formulation 3 of C from section I.8:

$$C = \{x: x = \sum_{n=1}^{\infty} \frac{d_n}{3^n} \text{ where } d_n = 0 \text{ or } 2\}.$$

Let $x \in C$. Then,

$$x = \sum_{n=1}^{\infty} \frac{d_n}{3^n} \quad \text{(tertiary expansion with 0's and 2's)}$$

$$f_1(x) = \frac{x}{3} = \sum_{n=1}^{\infty} \frac{d_n}{3^{n+1}} = \sum_{n=1}^{\infty} \frac{d_{n-1}}{3^n} \quad \text{where } d_0 = 0.$$

Thus, f_1 takes points in C to other points in C.

$$f_2(x) = \frac{x}{3} + \frac{2}{3} = \sum_{n=1}^{\infty} \frac{d_{n-1}}{3^n} + \frac{2}{3} \text{ ,where } d_0 = 0,$$

$$= \sum_{n=1}^{\infty} \frac{d_{n-1}}{3^n} \quad \text{where } d_0 = 2.$$

So, f_2 also takes points in C to the other points in C. Thus, C is a fixed (set) point of f.

How does one find the dimension of a fixed point of an IFS? We know the dimension of the Cantor Set is log 2/log 3, but in general the functions in the IFS must hold the key.

Definition. The ratio of a contraction map on R^n is the infimum of all r for which $d(f(x), f(y)) \leq rd(x,y)$ still holds for all $x, y \in R^n$.

Thus, in example 12, the ratios of f_1 and f_2 are both $\frac{1}{3}$.

Fact. Under certain conditions on the IFS given by $f = \{f_i : i = 1, \ldots, m\}$, if the ratio of f_i is r_i, then the dimension of the fixed point (set) of f is the unique number s such that $\sum_{i=1}^{m} r_i^s = 1$.

The conditions necessary for this fact are too technical to discuss. But, among other things, each f_i must be a similitude (preserves shape, but not size). Additionally, there cannot be too much overlap in a construction (following the process of section I.7).

Example 13. Calculate the Hausdorff dimension of the Cantor Set by using the fact on the previous page.

Since $r_1 = r_2 = \frac{1}{3}$, we need to solve $(\frac{1}{3})^s + (\frac{1}{3})^s = 1$ for s.

$2(\frac{1}{3})^s = 1$, so $s = \log(1/2)/\log(1/3)$, implying that $s = \log 2/\log 3$.

Exercises.

1. Prove that the topology generated by a metric space really is a topology.

2. Prove that the general Euclidean metric really is a metric.

3. In section I.8, let E_n be the n^{th} set in the recursive process which generates the Cantor set in formulation 1. (Alternately, E_n is the n^{th} iteration of f on [0,1] in example 12 of this section.) Find $d(E_n, [0,1])$ where d is the Hausdorff metric on compact subsets of R.

4. Let f be the IFS in example 12. Find $\lim_{n\to\infty} f^n([-1,0])$.

5. Let $f = \{f_1, f_2, f_3\}$ where each $f_i: R^2 \to R^2$, is given in polar coordinates by:

$$f_1(r,\theta) = (\frac{r}{2}, \theta + \frac{\pi}{6})$$

$$f_2(r,\theta) = (\frac{r+1}{2}, \theta + \frac{2\pi}{3})$$

$$f_3(r,\theta) = (\frac{1-r}{2}, \theta - \frac{\pi}{6}).$$

Calculate the Hausdorff dimension of the fixed point of f. Find f(A) where $A = \{(r,\theta): r \leq 1\}$.

6. Prove that a contraction mapping is continuous and has a unique fixed point.

I.12 Further Examples

It turns out that many useful IFS are collections of "affine" functions.

Definition. A function, $f: R^n \to R^n$, is affine if and only if there is an n x n matrix, A, and a n x 1 vector, $\bar{b}$ so that if $\bar{x} \in R^n$, ($\bar{x}$ is n x 1), then $f(\bar{x}) = A\bar{x} + \bar{b}$.

The functions in the IFS must still be contraction maps, but affine functions usually suffice to form useful pictures (compact fixed sets).

Example 1. Let $f = \{f_1, f_2, f_3\}$ where each $f_i: R^2 \to R^2$ and

$$f_1\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix},$$

$$f_2\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 1 \\ 50 \end{bmatrix}, \text{ and}$$

$$f_3\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 25 \\ 50 \end{bmatrix}.$$

This IFS has the "Sierpinski Triangle" as its fixed set (point). (Figure 6, page 57.)

Remember that the fixed set, P, is the $\lim_{n \to \infty} f^n(A)$, where A is any nonempty subset of R^2. The easiest A to choose would be the set containing just the origin;

$$A = \left\{ \begin{bmatrix} 0 \\ 0 \end{bmatrix} \right\}.$$

Then:

$f_1(A) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, $f_2(A) = \begin{bmatrix} 1 \\ 50 \end{bmatrix}$, and $f_3(A) = \begin{bmatrix} 25 \\ 50 \end{bmatrix}$.

So, $f(A) = \{\begin{bmatrix}1\\1\end{bmatrix}, \begin{bmatrix}1\\50\end{bmatrix}, \begin{bmatrix}25\\50\end{bmatrix}\}$.

$f^2(A) = \{\begin{bmatrix}1.5\\1.5\end{bmatrix}, \begin{bmatrix}1.5\\26\end{bmatrix}, \begin{bmatrix}13.5\\26\end{bmatrix}, \begin{bmatrix}1.5\\50.5\end{bmatrix}, \begin{bmatrix}1.5\\75\end{bmatrix}, \begin{bmatrix}13.5\\75\end{bmatrix}, \begin{bmatrix}25.5\\50.5\end{bmatrix}, \begin{bmatrix}25.5\\75\end{bmatrix}, \begin{bmatrix}37.5\\75\end{bmatrix}\}$, etc.

Implementing this algorithm on a computer is memory intensive. It is necessary to store $f^k(A)$, reference all its elements, and generate $f^{k+1}(A)$. To generate a picture on a typical computer monitor would require two arrays, each capable of holding 640 x 350 bits. And this is only for functions in R^2.

Here is another algorithm. Unfortunately, it is extremely slow and redundant, but it does require minimal memory.

Theorem. If $f = \{f_i: i = 1, \ldots, m\}$ is an IFS, then the fixed set (point) of f, P, is the collection of all fixed points of all finite compositions of the f_i's.

Example 2. Another way of constructing the fixed set of example 1.

We first find the fixed points of f_1, f_2, and f_3. (Note: the fixed point of $A\bar{x} + \bar{b}$ is $(I - A)^{-1}\bar{b}$ if it exists, where I is the identity matrix.)

The fixed point of f_1 is:

$$\left(\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix} - \begin{bmatrix}0.5 & 0\\0 & 0.5\end{bmatrix}\right)^{-1} \begin{bmatrix}1\\1\end{bmatrix}$$

$$= \begin{bmatrix}2 & 0\\0 & 2\end{bmatrix}\begin{bmatrix}1\\1\end{bmatrix} = \begin{bmatrix}2\\2\end{bmatrix}.$$

Similary, the fixed points of f_2 and f_3 are: $\begin{bmatrix}2\\100\end{bmatrix}$ and $\begin{bmatrix}50\\100\end{bmatrix}$, respectively. So, $\begin{bmatrix}2\\2\end{bmatrix}$, $\begin{bmatrix}2\\100\end{bmatrix}$, and $\begin{bmatrix}50\\100\end{bmatrix}$ are points in the fixed set of f, not just approximating points.

The theorem says we need <u>all</u> fixed points of <u>all</u> compositions of the f_i's. We just computed fixed points of compositions of order 1. I.e, f_1, f_2, f_3 are themselves trivial compositions.

Next we'll calculate compositions of order 2:

These are:

$f_1 \circ f_1$, $f_1 \circ f_2$, $f_1 \circ f_3$, $f_2 \circ f_1$, $f_2 \circ f_2$, $f_2 \circ f_3$, $f_3 \circ f_1$

$f_3 \circ f_2$, and $f_3 \circ f_3$.

Since composition of functions is <u>not necessarily commutative</u>, all nine need to be considered. To compose two affine functions: $f_1 = A\bar{x} + \bar{b}$ and $f_2 = C\bar{x} + \bar{d}$,

$$f_1 \circ f_2(\bar{x}) = A(C\bar{x} + \bar{d}) + \bar{b}$$

$$= AC\bar{x} + A\bar{d} + \bar{b}.$$

In our example, all the matrices are the same, so each matrix of the order 2 composition will be $\begin{bmatrix} .25 & 0 \\ 0 & .25 \end{bmatrix}$. However, there are nine possibilities for the vector part: $\begin{bmatrix} .5 & 0 \\ 0 & .5 \end{bmatrix}\bar{d} + \bar{b}$ where $\bar{d}$ and $\bar{b}$ can be $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$, $\begin{bmatrix} 1 \\ 50 \end{bmatrix}$, or $\begin{bmatrix} 25 \\ 50 \end{bmatrix}$. So, we get vector parts of : $\begin{bmatrix} 1.5 \\ 1.5 \end{bmatrix}$, $\begin{bmatrix} 1.5 \\ 26 \end{bmatrix}$, $\begin{bmatrix} 13.5 \\ 26 \end{bmatrix}$, $\begin{bmatrix} 1.5 \\ 50.5 \end{bmatrix}$, $\begin{bmatrix} 1.5 \\ 75 \end{bmatrix}$, $\begin{bmatrix} 13.5 \\ 75 \end{bmatrix}$, $\begin{bmatrix} 25.5 \\ 50.5 \end{bmatrix}$, $\begin{bmatrix} 25.5 \\ 75 \end{bmatrix}$, and $\begin{bmatrix} 37.5 \\ 75 \end{bmatrix}$.

We need to calculate the 9 fixed points of these 9 compositions. Since each composition is still an affine map, we use the same technique as on the order 1 compositions to obtain fixed points:

$\begin{bmatrix} 2 \\ 2 \end{bmatrix}$, $\begin{bmatrix} 2 \\ 34.7 \end{bmatrix}$, $\begin{bmatrix} 18 \\ 34.7 \end{bmatrix}$, $\begin{bmatrix} 2 \\ 67.3 \end{bmatrix}$, $\begin{bmatrix} 2 \\ 100 \end{bmatrix}$, $\begin{bmatrix} 18 \\ 100 \end{bmatrix}$, $\begin{bmatrix} 40 \\ 67.3 \end{bmatrix}$, $\begin{bmatrix} 40 \\ 100 \end{bmatrix}$, and $\begin{bmatrix} 50 \\ 100 \end{bmatrix}$.

Note that $\begin{bmatrix} 2 \\ 2 \end{bmatrix}$ and $\begin{bmatrix} 50 \\ 100 \end{bmatrix}$ were already known. Even though we calculated 3 + 9 fixed points, we only have 10 distinct points in our fixed set.

Since the number of compositions of order n is 3^n, and there is much redundancy, it should be clear that this algorithm is slow. But, at least it requires minimal memory, (just for the original functions).

Example 3. Define an IFS as $f = \{f_i : i = 1, \ldots, s\}$ where $f_i : R^2 \to R^2$ as

$$f_i(\bar{x}) = A_i\bar{x} + \bar{b}_i \text{ and}$$

$$A_1 = \begin{bmatrix} 0 & 0 \\ 0 & .18 \end{bmatrix}, \quad A_2 = \begin{bmatrix} .85 & 0 \\ 0 & .85 \end{bmatrix},$$

$$A_3 = \begin{bmatrix} -.85 & 0 \\ 0 & .85 \end{bmatrix}, \quad A_4 = \begin{bmatrix} .2 & .2 \\ .2 & .2 \end{bmatrix}, \text{ and}$$

$$A_5 = \begin{bmatrix} -.2 & .2 \\ .2 & .2 \end{bmatrix}.$$

$$\bar{b}_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \bar{b}_2 = \bar{b}_3 = \begin{bmatrix} 0 \\ 1.6 \end{bmatrix}, \text{ and } \bar{b}_4 = \bar{b}_5 = \begin{bmatrix} 0 \\ 0.8 \end{bmatrix}.$$

This IFS will generate a fern branch in R^2. It is possible to modify this IFS to generate a (curved) fern branch in R^3...

$$A_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & .18 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad A_2 = \begin{bmatrix} .85 & 0 & 0 \\ 0 & .85 & .1 \\ 0 & -.1 & .85 \end{bmatrix}$$

$$A_3 = \begin{bmatrix} -.85 & 0 & 0 \\ 0 & .85 & .1 \\ 0 & -.1 & .85 \end{bmatrix}, \quad A_4 = \begin{bmatrix} .2 & .2 & 0 \\ .2 & .2 & 0 \\ 0 & 0 & .3 \end{bmatrix}, \text{ and}$$

$$A_5 = \begin{bmatrix} -.2 & .2 & 0 \\ .2 & .2 & 0 \\ 0 & 0 & .3 \end{bmatrix}.$$

$$\bar{b}_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad \bar{b}_2 = \bar{b}_3 = \begin{bmatrix} 0 \\ 1.6 \\ 0 \end{bmatrix}, \text{ and}$$

$$\bar{b}_4 = \bar{b}_5 = \begin{bmatrix} 0 \\ .8 \\ 0 \end{bmatrix}.$$

The best known algorithm for generating a fixed set of an IFS is the "Random Iteration Algorithm" or "Chaos Game". It is an extrapolation of the scheme presented in example 1. Instead of computing the complete set function at each iteration, we randomly choose a single contraction map from the IFS to iterate the last point only.

The RIA (Random Iteration Algorithm) relies on fixed probabilities associated with each contraction map in the IFS. The probabilities are chosen to describe the area-relationship of each map to the whole.

Area-relationships of affine maps are easily calculated by determinants. For example, if $f(\bar{x}) = A\bar{x} + \bar{b}$ then f will map the unit square (in R^2) with vertices: (0,0), (1,0), (1,1), and (0,1) to a parallelegram with area equal to the determinant of A.

Thus, if an IFS is given by $\{f_i : i = 1, \ldots, n\}$ and each $f_i(\bar{x}) = A_i\bar{x} + \bar{b}_i$, then we will assign a probability, p_i, to each f_i via the formula:

$$p_i = \frac{\det A_i}{\sum_{i=1}^{n} \det A_i} .$$

This will insure that those contraction maps dealing with "a lot of area" have a higher probability. If $\det A_j = 0$, then assign a small probability like 0.01 to f_j.

The RIA can be described as follows:

1. Given an "IFS with probabilities": $\{f_1, p_1; f_2, p_2; \ldots; f_n, p_n\}$ so that $\sum_{i=1}^{n} p_i = 1$.

2. Pick a fixed point, x_{old}, of one of the f_i's.

3. Generate a uniform random number, r, between zero and one.

4. Choose k so that $\sum_{i=1}^{k-1} p_i < r \leq \sum_{i=1}^{k} p_i$.

5. Let $x_{new} = f_k(x_{old})$ and graph x_{old}.

6. Let x_{old} be replaced by x_{new} and repeat steps 3 through 6.

The RIA is very efficient. All the figures in this section were created with the RIA.

FACT: Small changes in the parameters (entries of the matrices and vectors) of an IFS will cause small changes in the fixed set (picture).

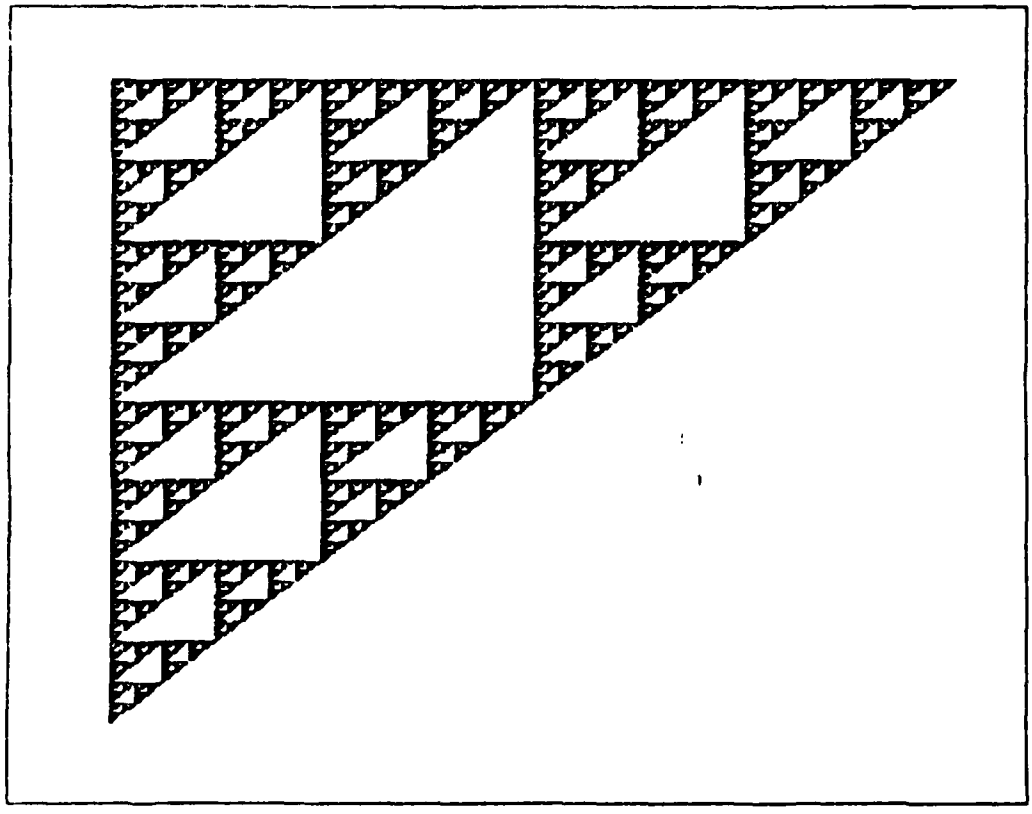


Figure 6

"This Sierpinski Triangle is the fixed set of three contraction maps: $f_i(\bar{x}) = A_i\bar{x} + b_i$, $i = 1, 2, 3$ where

$$A_1 = A_2 = A_3 = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix}, \quad \bar{b}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix},$$

$$\bar{b}_2 = \begin{bmatrix} 1 \\ 50 \end{bmatrix}, \text{ and } \bar{b}_3 = \begin{bmatrix} 50 \\ 50 \end{bmatrix}."$$

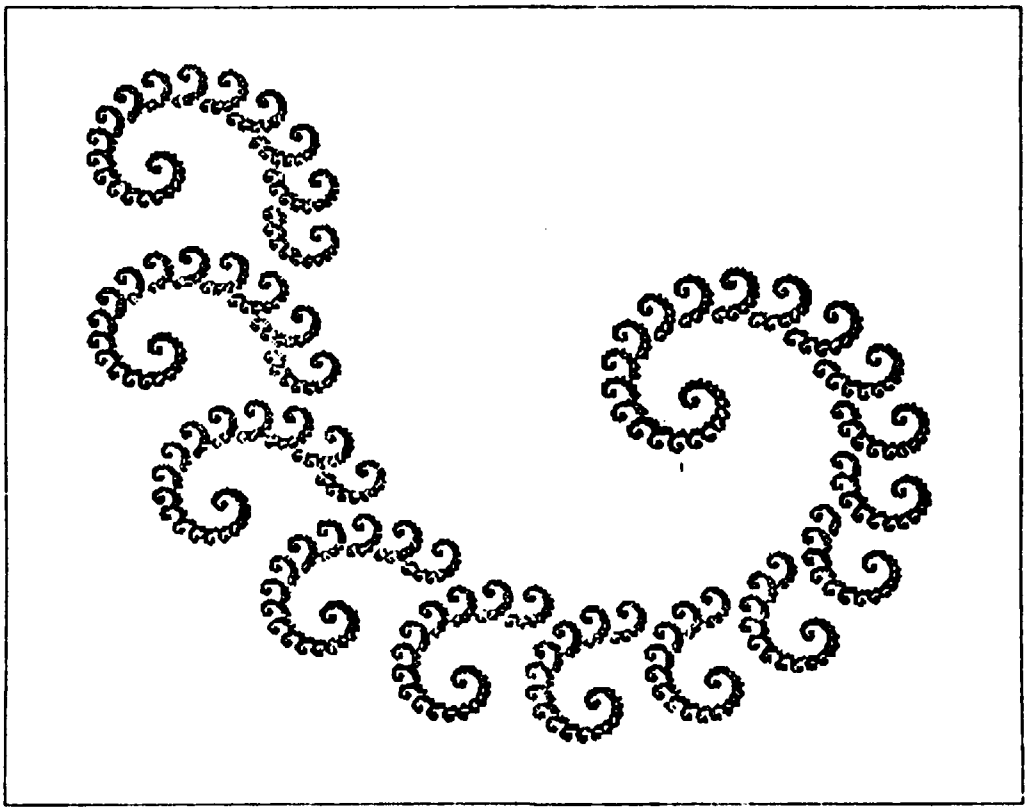


Figure 7

"This spiral was created using two contraction maps with matrices and vectors given by

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \text{ and } \begin{bmatrix} e \\ f \end{bmatrix}$$

where

	a	b	c	d	e	f
f_1	.85	-.31	.31	.85	1	-10
f_2	-.3	0	0	-.3	10	-1

."

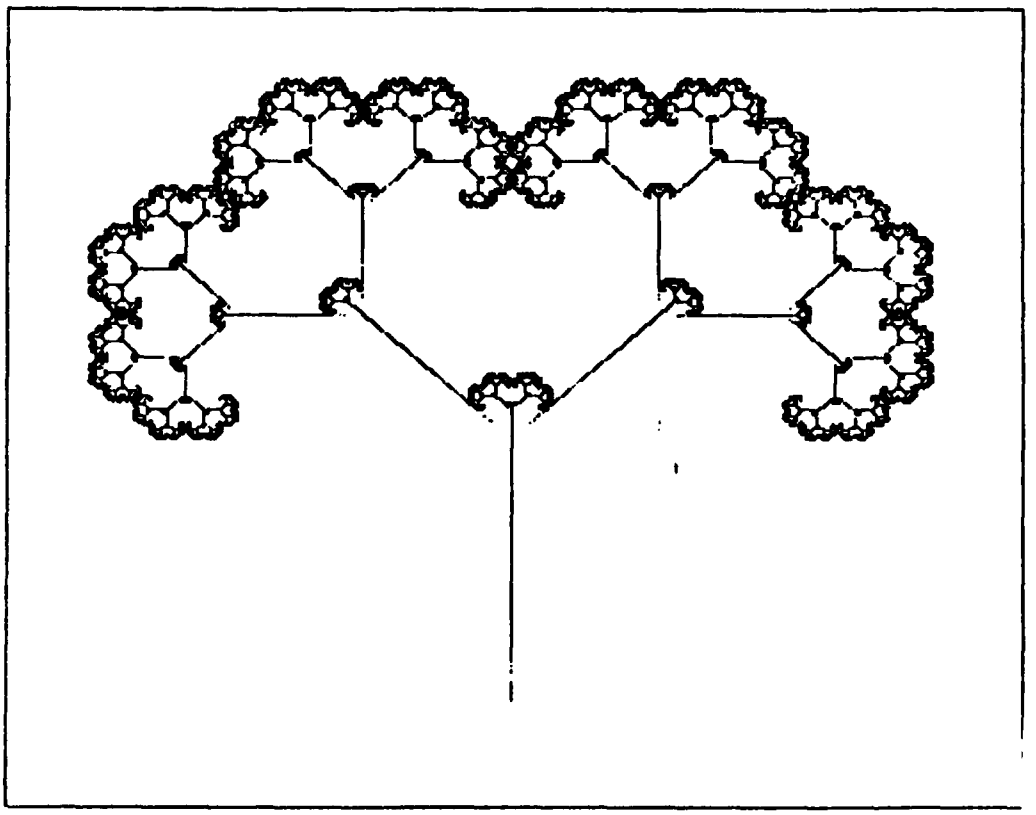


Figure 8

"Continuing the notation from Figure 7, this figure is the fixed set of the IFS given by:

	a	b	c	d	e	f
f_1	0	0	0	.5	0	0
f_2	.42	-.42	.42	.42	0	.2
f_3	.42	.42	-.42	.42	0	.2
f_4	.1	0	0	.1	0	.2

."

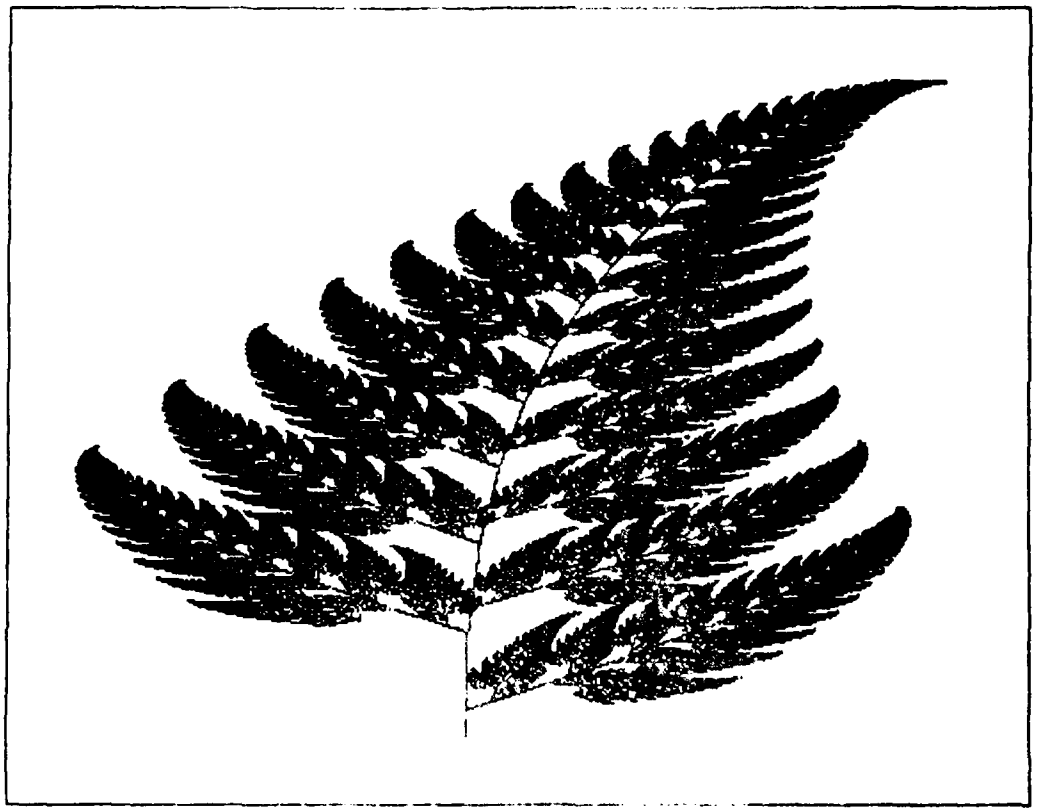


Fig. 6

"This fern branch also appears on page [illegible] IFS is:

	a	b	c	d	e	f
f_1	0	0	0	[illegible]	[illegible]	[illegible]
f_2	.85	.04	[illegible]	[illegible]	[illegible]	[illegible]
f_3	.2	-.26	[illegible]	[illegible]	[illegible]	[illegible]
f_4	-.15	.28	[illegible]	.	[illegible]	.44 ."

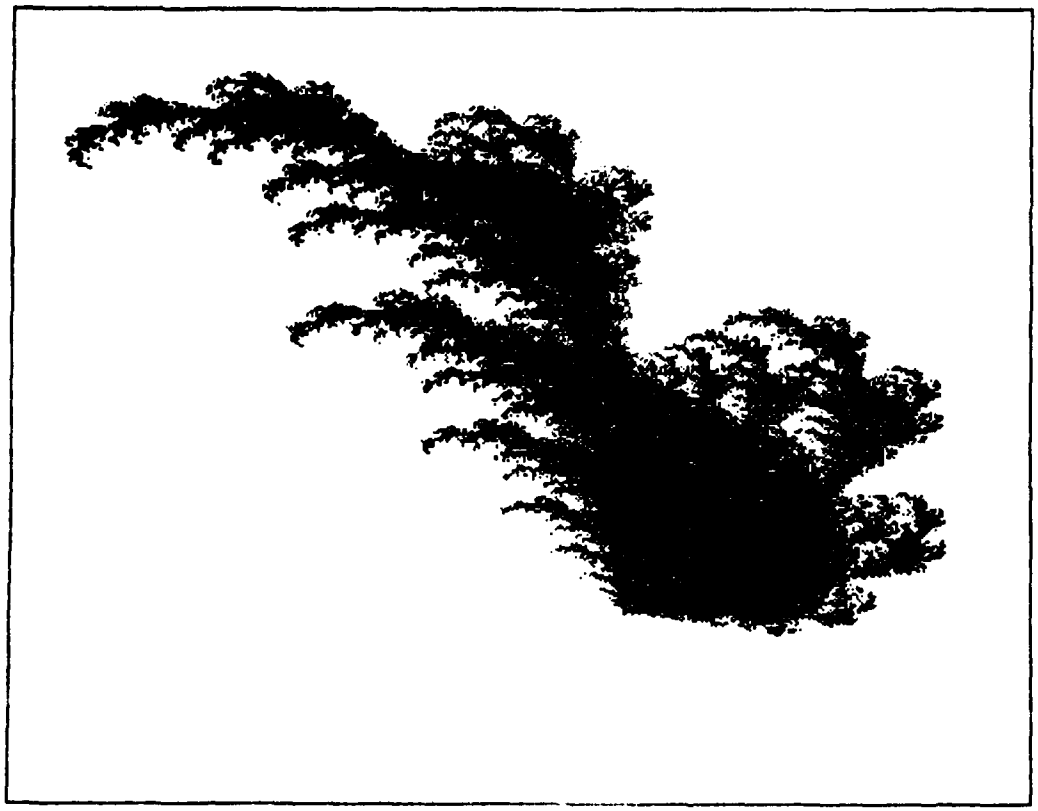
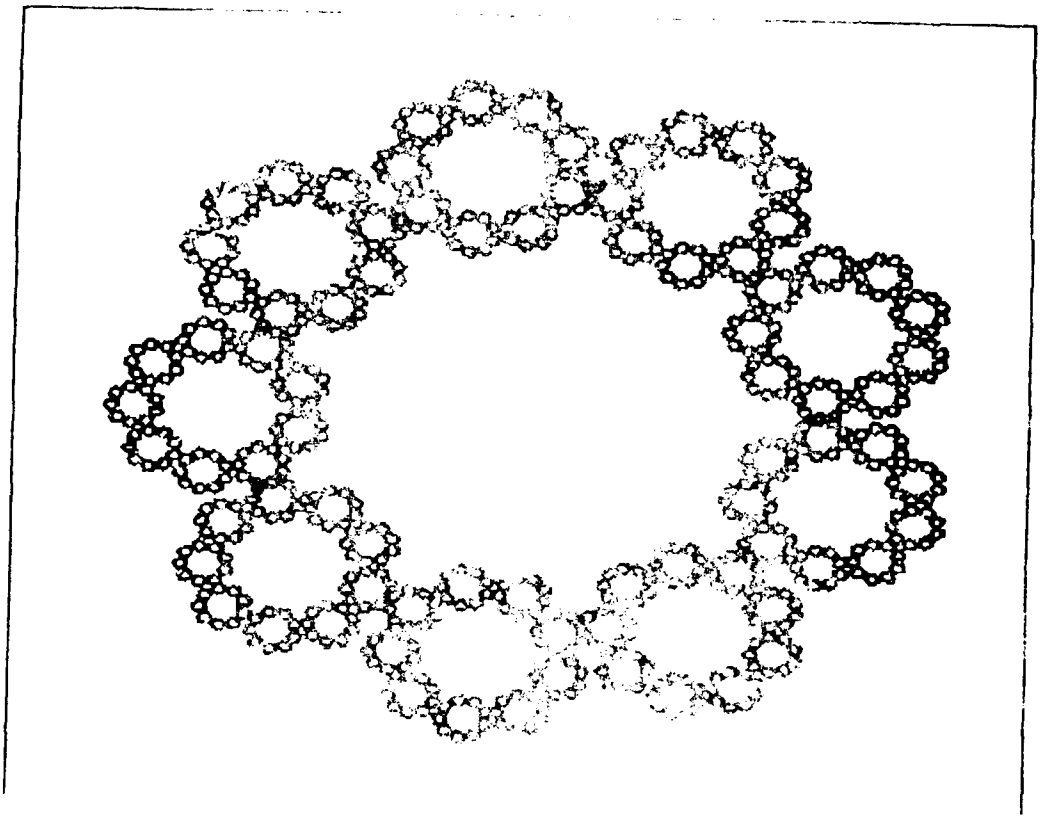


Figure 10

"The IFS for this figure is:

	a	b	c	d	e	f
f_1	.6	0	0	.6	.18	.36
f_2	.6	0	0	.6	.18	.12
f_3	.24	.3	-.3	.4	.27	.36
f_4	.4	-.3	.3	.4	.27	.9

."



"The IFS f

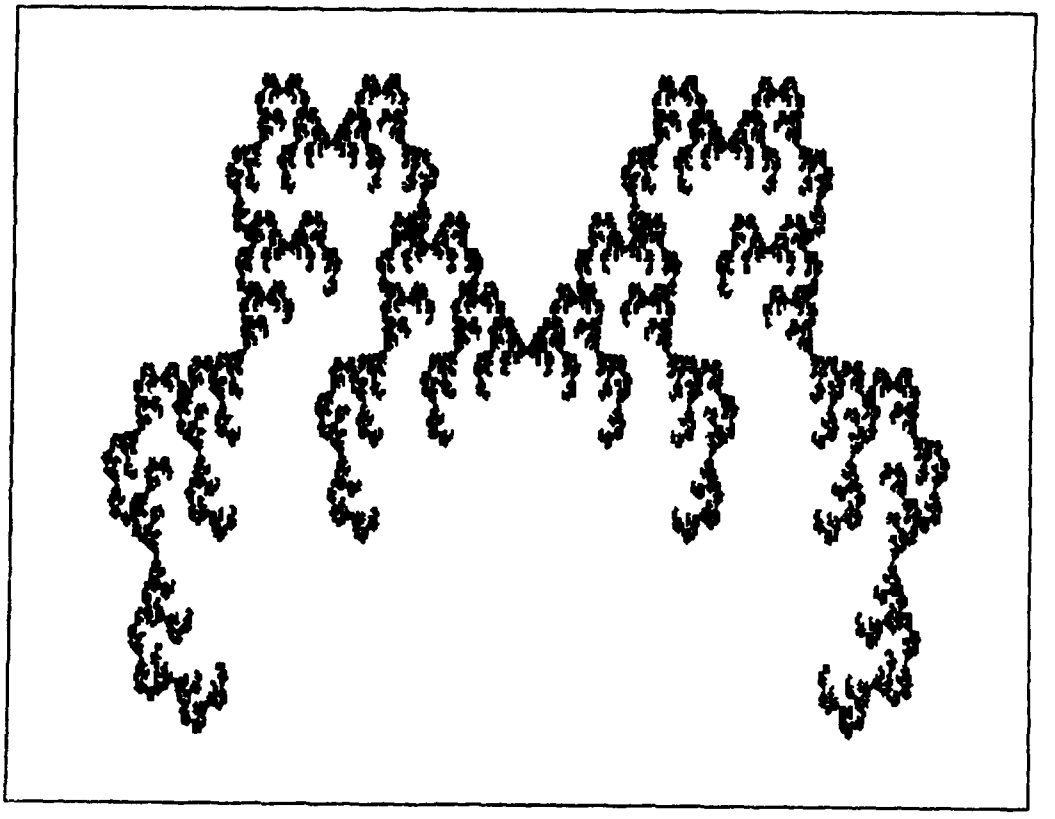


Figure 12

"This figure's IFS is:

	a	b	c	d	e	f
f_1	.3	-.4	.4	.3	1	0
f_2	.5	0	0	.5	0	0
f_3	.3	.4	-.4	.3	-1	0

."

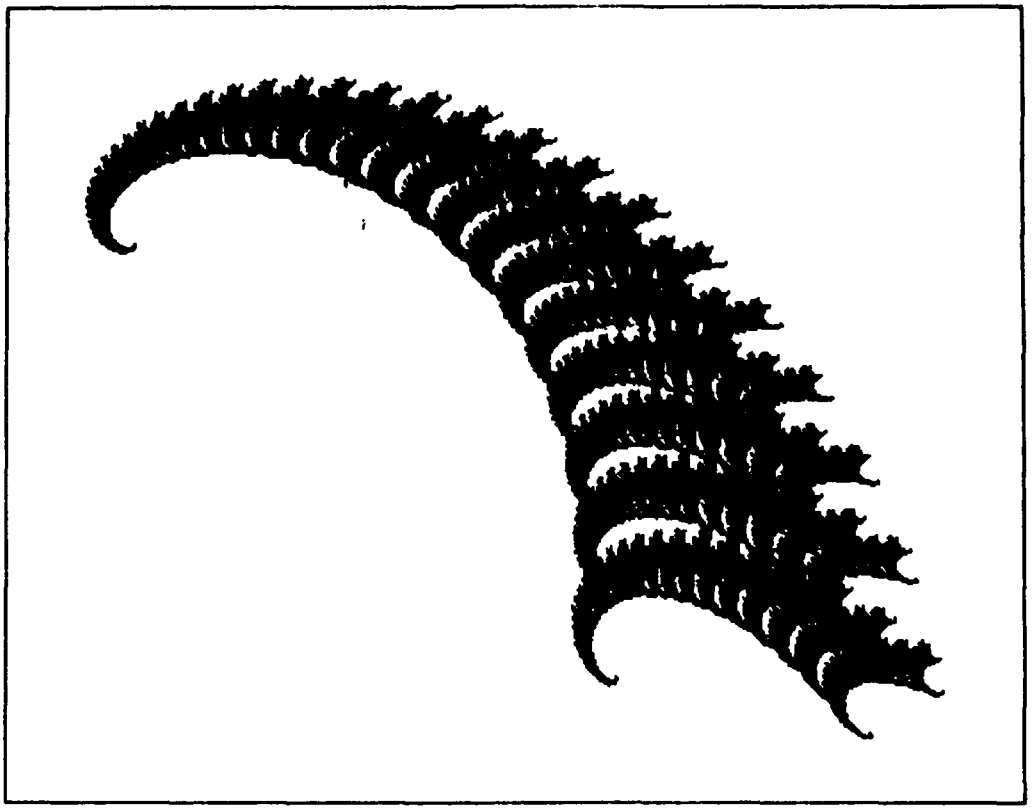


Figure 13

"This dragon's tail was created with an IFS of:

	a	b	c	d	e	f
f_1	.2	-.3	.3	.4	10	-1
f_2	.9	-.1	.1	.9	-1	10

."

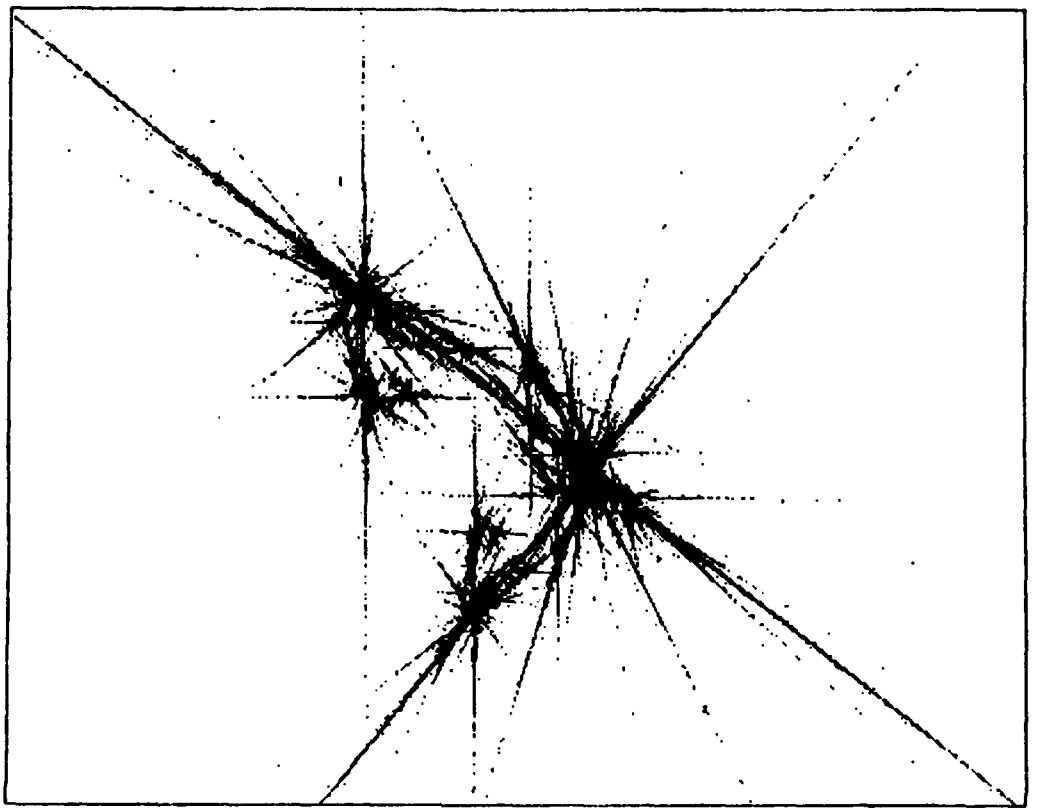


Figure 14

"This star-burst fractal was created with an IFS of:

	a	b	c	d	e	f
f_1	-.707	.707	.707	0	10	1
f_2	.5	0	0	-.8	10	1
f_3	0	0	.5	-.5	10	0

."

Exercises.

1. What is the dimension of the Sierpinski Triangle?
2. Recall that the matrix $\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$ will rotate points in R^2, θ radians, counterclockwise about the orgin. Write $\begin{bmatrix} .85 & .1 \\ -.1 & .85 \end{bmatrix}$ as $r\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$. What is r and θ. Note that r is the contraction ratio.
3. What is the approximate dimension of the fern in R^3 described in the second half of example 3? (Hint: When dealing with a function which doesn't have an easy contraction ratio, bound it by 2 functions which do. This will give bounds on the dimension when using the fact in section I.11.)
4. If $f(\bar{x}) = A\bar{x} + \bar{b}$, under what conditions will I-A be an invertable matrix, so that the fixed point of f may be found? (Hint: Consider the eigenvalues of A.)
5. It can be shown that if A is a 2 x 2 matrix and $f(x) = A\bar{x} + \bar{b}$, then f is a contraction map if and only f:

$$\frac{\|A\|^2 + \sqrt{\|A\|^4 - 4|A|^2}}{2} < 1.$$

In fact, the ratio of f is the quantity on the left-hand-side where: $\|A\|^2 = a^2 + b^2 + c^2 + d^2$ and $|A|^2 = (\det A)^2$. Find the dimension of the fractal in figure 14. What is the ratio of a rotation map?

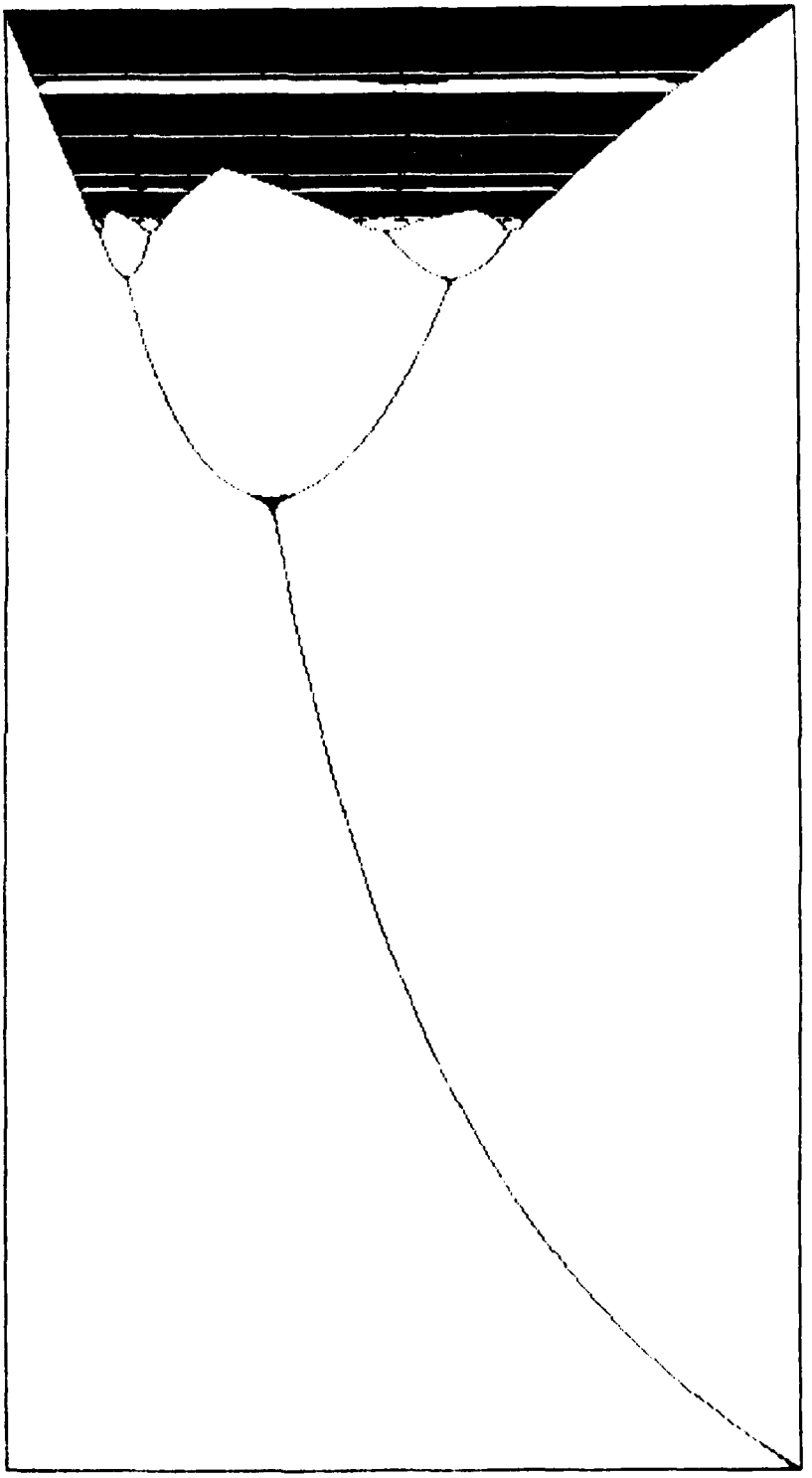


Figure 15

II.1 Introduction

The study of chaos is better described as a study of the paths to chaos. Typically, a physical system--like a driven pendulum--will undergo predictable regular behavior through a continuous change of a parameter (in the motor), until all at once, irregularity is observed. Or, a column of smoke from a cigarette will rise in easy patterns, until at some height it breaks up and becomes turbulent. Or, the fluid flow in a blender seems to follow regular swirls, until at some speed, it looks random.

Obviously, there are magnitudes of irregularity, and chaos is not always an apt term for some of the less dramatic behavior of these systems. But the connotation that the word chaos provides is appropriate: there is an order of complexity in the system which makes it unpredictable. This complexity is usually a symptom of a nonlinear system. Linear systems have highly regular behavior through all changes in their parameters. But a nonlinear system has an amazing potential for strange behavior.

Some thought on observations of physical systems leads us to suspect noise, and at very small scales, quantum effects as the culprits behind unpredictability. Noise and Heisenberg's uncertainty principle are not explained away or even encompassed by chaos theory. They are additional limits to observations. Chaos provides randomness from seemingly exact mathematical models. Noise and quantum effects provide additional unpredictability to the exact model. This report will no address the effects of noise and quantum principles.

First, we will discuss what a nonlinear function can do when applied repeatedly to a point (iterated). This is the study of orbits, and it leads to invariant sets and attractors, all of which have physical significance. We will tie this to fractals by noting that many natural invariant sets and attractors are also fractals.

Perhaps the most astonishing facts in chaos theory come from universality. Mitchell Feigenbaum has shown that many of the paths to chaos are essentially the same. When viewed from the parameter space, (the parameter whose change will bring unpredictability), the progression to chaos of a driven pendulum will be "the same" as that of a small cell of heated fluid. Two univeral constants, $\alpha = 2.50290787...$ and $\delta = 4.6692016...$, will show up in both of these systems. These discussions will focus on the "bifurcation diagram."

After universality, we will also present some basic complex number dynamics, to include the Mandelbrot Set, and finish with a discussion of randomness.

II.2 The Poincaré Map

One of the most useful spaces in which to describe physical phenomena is called phase space. It frequently involves twice the number of dimensions needed to describe the system itself because phase space usually contains information on both the position and velocity of the system. We are plotting dependent variables against each other. For example, x(t) and v(t). Phase space allows us to tell how a system evolves in time. The path (trajectory) in phase space shows how the system behaves. If the path is confined to a region, that says something about the system's nature.

If a pendulum is oscillating in a plane, then only the angle, θ, is necessary to describe the pendulum's position. So, it is a one-dimensional system. Rather than graph θ versus time, t, phase space will plot $\dot{\theta}$ versus θ (where $\dot{\theta} = d\theta/dt$).

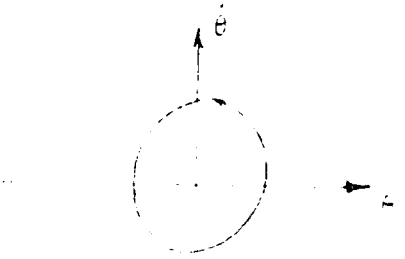
If the pendulum is not oscillating in a plane, (this is realistic due to the Coriolis effect), then two angles are necessary to describe its position. Thus, a graph of position versus time would require three dimensions, whereas the phase space is four dimensional.

Because phase space is usually of a dimension which precludes graphing, it is often useful to use a Poincaré section on a difficult trajectory. The trajectory in the first example will be close to a circle, parametrized by time (and thus, with direction). But, if the pendulum is driven, the trajectory need not close back on itself. One way of simplifying this is to intersect the trajectory with a curve (or line). This curve, with its points of intersection, (usually labeled with the time of intersection), is the Poincaré section.

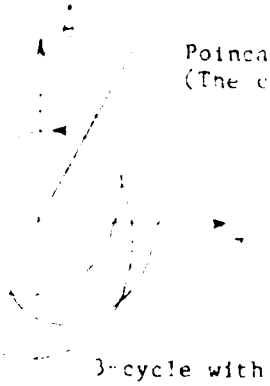
The Poincaré map is the function which gives a point of intersection from the trajectory and the Poincaré section. In the case of the driven pendulum, we will obtain a different Poincaré map (and Poincaré section) with each change in

Pendulum

Simple Harmonic Motion
(Position vs Time)



Simple Harmonic Motion
(Phase Space)

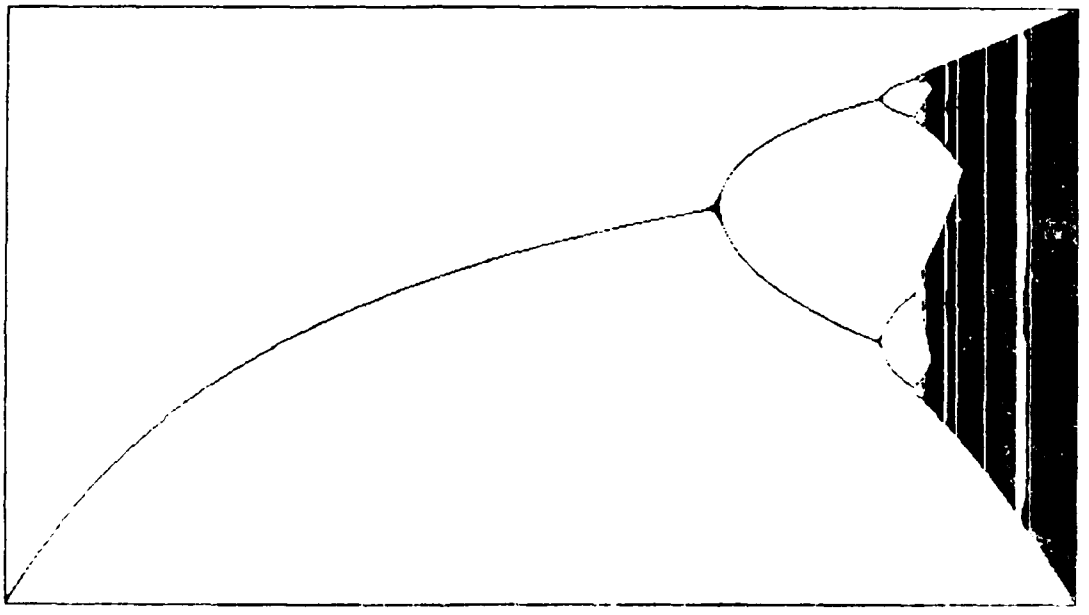


Poincaré Section
(The curve is at time: 0, P/2, and P)

3-cycle with period, P

Chaos
(No period)

the nonlinearity constant: that number which describes the driver. For very small values of this number, we will still obtain a 1-cycle (closed curve). Then at a fixed value, a 2-cycle will appear. For smaller and smaller changes in the nonlinearity constant, 4-cycles, 8-cycles, etc. will develop until at last, chaos will prevail. In this case, chaos simply means that there is no cycle, and so the motion is irregular. This progression to chaos is best seen through a bifurcation diagram. (See II.7).



Another version of the Poincaré map (also called the return map) is obtained by fixing a value for the nonlinearity parameter and fixing a Poincaré section. We then start the trajectory at an arbitrary point in phase space and number the consecutive intersections with the Poincaré section (instead of just looking at the limit trajectory and labeling the time of intersection). A function, f, is defined as $f(x_n) = x_{n+1}$, where x_n and x_{n+1} are consecutive

positions on the Poincaré section. By starting the trajectory at all points in phase space, we obtain a continuous function, f. It turns out that nearly all such Poincaré maps have the same generic shape: that of an inverted parabola, (although it usually is bounded since we usually limit ourselves to a bounded region in phase space).

Because of the way this Poincaré map is defined, the fundamental technique in studying it is the iteration of continous functions. If repeated applications of the function, f, bring us closer to a single point, then we are approaching a fixed point of f, and it corresponds to a 1-cycle in phase space.

Exercises.

1. An ideal pendulum (with planar motion) with no friction will regularly repeat any state it is started in. Thus, the limit trajectory is always a 1-cycle. If we measure theta from $-\pi$ to π, reason that the Poincare map (return map) will be: $f(x) = x$, for $-\pi < x < \pi$. (Use the θ-axis as the Poincaré section.)

2. Do the same as in problem 1 but use the $\dot{\theta}$-axis as the Poincaré section. What will the domain of the return map be?

II.3 Iterations of Continuous Functions: Orbits

Because the Poincaré map can reduce the dimension and complexity of a physical system's behavior, it is very useful in the study of chaos. The return map is usually a continuous function defined on an interval which is strictly increasing to a maximum and then decreasing. This generic characteristic of most Poincaré maps led mathematicians to study their behavior without reference to any physical system.

The most commonly used technique to analyze behavior of return maps involves repeated iterations of functional values. We start with a fixed value, x_0, and compute $f(x_0) = x_1$, $f(f(x_0)) = f(x_1) = x_2$, etc., to obtain the orbit of x_0. The orbit is a sequence of numbers, $\{x_n\}_{n=0}^{\infty}$, which can display many types of behavior:

1. Each x_n is either the same, or eventually the same. That is, $\{x_n\}$ is a constant sequence. Recalling that the Poincaré map is derived from a fixed Poincaré section and nonlinearity parameter, the significance of this is that the physical system is going through a 1-cycle.

2. Every m values of the orbit, $\{x_n\}$, repeat. This implies that the system is going through an m-cycle.

3. The orbit gets closer and closer to m repeating values. This system has an m-cycle as its limit trajectory, but the particular orbit chosen is not on that trajectory. It merely approaches it.

4. The orbit, $\{x_n\}$, is dense in the domain of the return map. Most definitions will include this as a criterion for chaos. We can imagine the trajectory of the physical system as randomly filling up its phase space: like a motorized spinning cue ball rebounding on a billiard table (without pockets).

5. The orbit seems to be random, as in 4, but is localized: that is, it never goes in some regions of the domain. The trajectory of this system is influenced by some attractor, quite possibly a fractal. (These are often called strange attractors, and will be discussed more in section II.5.)

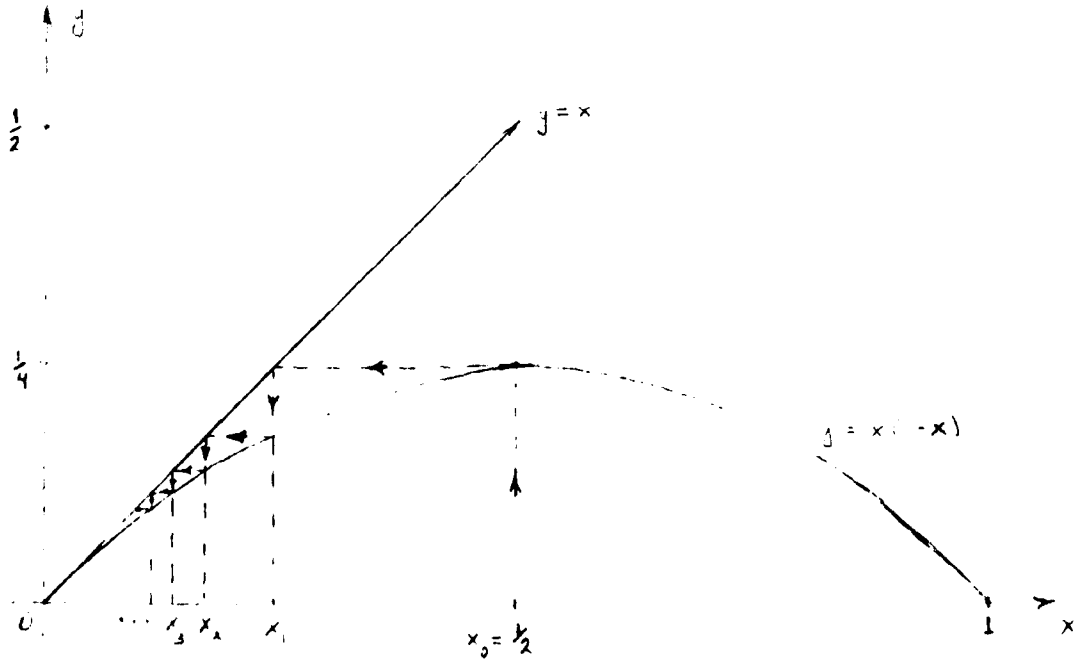
All of the behaviors listed above, while significant to the physical system, can be studied using simple mathematical tools.

A useful geometric trick is described in the following example:

Example 1.

Let $f(x) = x(1 - x)$, for $x \in [0,1]$. Find the orbit of $x_0 = \frac{1}{2}$.

The straightforward approach involves finding $f(\frac{1}{2})$, $f^2(\frac{1}{2}) = f(f(\frac{1}{2}))$, etc.: we would get $\frac{1}{2}$, $\frac{1}{4}$, $\frac{3}{16}$, $\frac{13}{256}$, ... The arithmetic gets tedious, and a calculator or computer is useful. But the qualitative behavior of the orbit of $\frac{1}{2}$ can be gleaned from the following geometric technique:



If we believe the diagram, it is obvious that the orbit is converging to zero. This technique works as follows: We start at x_0 and go up to the graph of f to find $f(x_0)$. This value is x_1 but is represented as a height on the y-axis. To represent x_1 on the x-axis, we go horizontally to the line, $y = x$, and then vertically (down) to the x-axis. The distance from zero to x_1 is exactly the height of $x_1 = f(x_0)$. We then repeat this process ad infinitum to observe the orbit, $\{x_n\}$. In this case, the x_n's obviously decrease to zero, a limit point. (This corresponds to a physical system approaching a 1-cycle, or losing energy to a state of no motion.)

Example 2.

Define $f_\lambda(x) = \lambda x(1 - x)$, for $x \in [0, 1]$. When $\lambda = 1$, the function of example 1 is obtained. λ is a nonlinearity parameter, and changes in λ can put us on the path to chaos. (See the exercises.)

Example 3.

Define $g_\lambda(x) = \begin{cases} \lambda x, & \text{if } 0 \leq x \leq \frac{1}{2} \\ \lambda(1 - x), & \text{if } \frac{1}{2} \leq x \leq 1. \end{cases}$

Again, λ is a nonlinearity parameter. The functions defined by g are called tent maps because of their shape. The functions defined by f are called logistic maps.

When $\lambda = 2$, g_2 can be analyzed with a numerical trick:

Write the numbers in [0, 1] in their binary expansion. Then note that since $g_2(x) = 2x$ or $g_2(x) = 2(1 - x)$, (and 2 = 10 base 2), calculations of g_2 are very easy. In the following explanation, all numbers with a point, ".", are base 2.

For example, if $x_0 = .01011$, then since the first digit is a zero, $x_0 \leq \frac{1}{2}$, so $g_2(x_0) = 2x_0 = .1011 = x_1$. Now, $x_1 \geq \frac{1}{2}$, so $g_2(x_1) = 2(1-x_1) = 2(.0101) = .101 = x_2$. Then, $g_2(x_2) = 2(1-x_2) = .11 = x_3$. $g_2(x_3) = 2(1-x_3) = .1 = x_4$, $g_2(x_4) = 1 = x_5$, and $g_2(x_5) = 0$. So, $x_6 = x_7 = \cdots = 0$. Thus, the orbit is: $\{.01011, .1011, .101, .11, .1, 1, 0, 0, \ldots\}$.

Note that the line $y = x$ will intersect g_2 at $x = 2(1-x)$; or $x = \frac{2}{3}$. $\frac{2}{3} = .\overline{10}$ base 2, so we can check that .101010 ... is really a fixed point of g_2: $g_2(\frac{2}{3}) = 2(1 - \frac{2}{3}) = \frac{2}{3}$. We can note a few things about the return map given by g_2. First, any x_0 which can be expressed as a finite decimal in base 2, will have an orbit which is eventually always zero. Zero and two-thirds are the only fixed points and the only points of g_2 which will have orbits that fix onto $\frac{2}{3}$ must end in $\overline{10}$ (base 2).

In fact, every rational number in [0,1] will have an orbit which eventually corresponds to an m-cycle, for some m. (The above remarks are detailed in the exercises.)

This has an interesting application to computer calculations with g_2. We might expect irrational orbits to behave chaotically; but, this cannot be simulated on a computer since any computer representation of an irrational number is, in fact, rational. If done correctly in base 2, then the computer will always end up at zero.

There is a technique which can glean more information from chaotic maps. It is called symbolic dynamics, and we recommend An Introduction to Chaotic Dynamical Systems by Robert Devaney.

Exercises.

1. In example 1, find the orbit of 0. What is the orbit of 1? And $\frac{1}{3}$?
2. In example 2, find all the fixed points of f_λ when $0 < \lambda \leq 4$.
3. For $\lambda = 2$, find the orbits of 0, 1, and $\frac{1}{3}$ under $f_\lambda(x) = \lambda x(1 - x)$.
4. For f_λ as above, what value of λ will force f_λ to intersect the line, $y = 1$, at $x = \frac{1}{3}$ and $x = \frac{2}{3}$.
5. Same as problem 4 with f_λ replaced by g_λ (the tent map) of example 3.
6. In problems 4 and 5, for the given value of λ, if $\frac{1}{3} < x < \frac{2}{3}$ then $f_\lambda(x)$ and $g_\lambda(x)$ are not in the interval [0,1]. Find the values of x such that $f_\lambda^n(x)$ is still in [0,1] for every value of n. Will these values of x change when we examine g_λ?
7. In example 3, find the orbit of .010111 base 2 under g_2.
8. Find the orbit of $\frac{1}{10}$ under g_2.
9. Prove that if x_0 has a finite representation, base 2, then $g_2^n(x_0)$ is eventually zero.
10. Prove that if x_0 is rational with an infinite base 2 representation that ends in repeating "10", then $g_2^n(x_0)$ is eventually $\frac{2}{3}$.
11. Prove that if x_0 is rational, then there exists a positive integer, m, so that the orbit of x_0 under g_2 will eventually repeat exactly m values.
12. Show that there is an x_0 (irrational) such that the orbit of x_0 is dense in [0,1] under g_2.

II.4 Periodic Points

Periodic points are points whose orbits (under the return map) repeat a finite sequence.

Definition 1. The point, x_0, is periodic under f if and only if there is a positive integer, n, so that $f^n(x_0) = x_0$. If n is the least such integer, then n is called the prime period of x_0.

Thus, a fixed point has prime period one and corresponds to a 1-cycle in phase space. Similarly, if x_0 has prime period m, then there is an m-cycle in phase space which will intersect the Poincaré Section at x_0 and then at m - 1 other values before returning again to x_0.

There are two fundamental questions pertaining to periodic points:

1. Is a given periodic point an attractor?
2. How many periodic points are there? (And where?)

Definition 2. A periodic point, x_0, of period m of f is an attractor if and only if there is an an $\varepsilon > 0$ so that if $|x - x_0| < \varepsilon$ then $\lim_{n\to\infty} f^{mn}(x) = x_0$.

The periodic point, x_0, is a repellor if and only if there is an $\varepsilon > 0$ so that if $0 < |x - x_0| < \varepsilon$ then there is a k such that $|f^{km}(x) - x_0| \geq \varepsilon$.

Attractors are also called sinks if they are fixed points, just as repellors are called sources.

Fortunately, it is usually easy to determine whether a given periodic point is an attractor or repellor:

Theorem 1. If x_0 is a fixed point of f and $|f'(x_0)| < 1$, then x_0 is an attractor. If $|f'(x_0)| > 1$, then x_0 is a repellor.

Proof: Assume $|f'(x_0)| < 1$. Then there is an $\varepsilon > 0$ so that if $|x - x_0| < \varepsilon$,

$$\left|\frac{f(x) - f(x_0)}{x - x_0}\right| \leq r \text{ where } r < 1.$$

Thus, for all x in the interval $(x_0 - \varepsilon, x_0 + \varepsilon)$, $|f(x) - f(x_0)| \leq r\,|x - x_0|$, which shows that f is a contraction mapping on $(x_0 - \varepsilon, x_0 + \varepsilon)$ with fixed point, x_0. So, if $|x - x_0| < \varepsilon$, then $\lim_{n\to\infty} f(x) = x_0$.

Assume $|f'(x_0)| > 1$. Then there is a $\delta > 0$ so that if $|x - x_0| < \delta$,

$$\left|\frac{f(x) - f(x_0)}{x - x_0}\right| \geq q \text{ where } q > 1.$$

Hence, on $(x_0 - \delta, x_0 + \delta)$, $|f(x) - f(x_0)| \geq q\,|x - x_0|$. So, given x such that $|x - x_0| < \delta$, f(x) is farther from x_0 than x is; $(q > 1)$. Thus, there is k and δ so that $|f^k(x) - x_0| \geq \delta$, completing the proof.

It is easy to extend theorem 1 to general periodic points:

Theorem 2. If x_0 has prime period m under f, then x_0 is attracting if $\left|\frac{d}{dx} f^m(x_0)\right| < 1$ and repelling if $\left|\frac{d}{dx} f^m(x_0)\right| > 1$. Moreover, if x'_0 is any point in the orbit of x_0, then $\frac{d}{dx} f^m(x'_0) = \frac{d}{dx} f^m(x_0)$, implying that the orbit of x_0 shares the properties of x_0 itself when $\left|\frac{d}{dx} f^m(x_0)\right| \neq 1$.

Proof: First assume $\left|\frac{d}{dx} f^m(x_0)\right| < 1$. Then x_0 is an attracting fixed point of f^m.

So, we need only show that if x'_0 is in the orbit of x_0, then $\frac{d}{dx} f^m(x_0) = \frac{d}{dx} f^m(x'_0)\ldots$

But, $\frac{d}{dx} f^m(x_0) = \prod_{i=0}^{m-1} f'(x_i)$ by the chain rule, where $x_i = f(x_{i-1})$ for $i = 1, \ldots, m - 1$. Thus, the derivative of f^m at x_0 is the same as that at any other point in the orbit of x_0, completing the proof.

Theorem 2 can be used with the definition of an attracting periodic point to show that if $\left|\frac{d}{dx}f^m(x_0)\right| < 1$, (where x_0 has prime period m under f), then there is an $\varepsilon > 0$ such that when $|x - x_0| < \varepsilon$, $\lim_{n\to\infty} |f^n(x) - f^n(x_0)| = 0$. Essentially, the entire orbit attracts x. This will be discussed more in the next section.

Definition 3. A periodic point, x_0, of prime period m under f is called hyperbolic if and only if $\frac{d}{df}f^m(x_0) \neq 1$. Otherwise it is nonhyperbolic. (Branch points in a bifurcation diagram are always nonhyperbolic. See section II.7.)

Example 1. Let $f_\lambda(x) = \lambda x(1 - x)$ be the logistic map introduced in the last section, for $\lambda > 0$, and $0 \leq x \leq 1$.

First, find all periodic points and classify them when $0 < \lambda < 1$.

It is easy to see that zero is the only periodic point of f_λ when $0 < \lambda < 1$. $f'_\lambda(0) = \lambda$, so zero is attracting when $\lambda < 1$.

When $\lambda = 1$, zero is nonhyperbolic, but still attracting (weakly attracting).

Next, we'll consider $\lambda > 1$.

Here, f_λ has another fixed point at $(\lambda - 1)/\lambda$. So f_λ has 2 fixed points. $f'\left(\frac{\lambda-1}{\lambda}\right) = 2 - \lambda$, showing that $(\lambda - 1)/\lambda$ is attracting and hyperbolic when $1 < \lambda < 3$ and repelling and hyperbolic for $\lambda > 3$. Geometric considerations show that $(\lambda - 1)/\lambda$ is weakly repelling (nonhyperbolic) for $\lambda = 3$.

The above completely describes the fixed point behavior of f_λ.

Next, we'll try to find points of prime period 2...

To do this, we'll look for fixed points of: $f_\lambda^2(x) = -\lambda^3x^4 + 2\lambda^3x^3 - (\lambda^2 + \lambda^3)x^2 + \lambda^2x$. Thus, we'll solve $f_\lambda^2(x) - x = 0$, which must have 0 and $(\lambda - 1)/\lambda$ as solutions ($\lambda > 1$) since they are fixed points of f_λ^2. Thus, we can reduce $f_\lambda^2(x) - x = 0$ to a quadratic equation:

$-\lambda^3x^2 + (\lambda^3 + \lambda^2)x - (\lambda^2 + \lambda) = 0$, or $\lambda^2x^2 - (\lambda^2 + \lambda)x + (\lambda + 1) = 0$. The discriminant of the latter quadratic is:

$$D = \lambda^2(\lambda + 1)(\lambda - 3).$$

Since we are assuming $\lambda > 1$, (in order to have $(\lambda - 1)/\lambda$ be a fixed point), we can see that D is negative until $\lambda \geq 3$. Thus, there are no periodic points of prime period 2 until $\lambda > 3$. (When $\lambda = 3$, the discriminant is zero and the old fixed point: $(\lambda - 1)/\lambda$, has multiplicity 3).

Note that the roots of $f_\lambda^2(x) - x$ are a continuous function of λ. So, at $\lambda = 3$ we have one repelling fixed point at zero (a root of multiplicity one), and a weakly repelling fixed point at 2/3, $(\lambda - 1)/\lambda$, which is a root of multiplicity 3. This root at 2/3, will split into 3 distinct roots for $x > 3$. One of them is $(\lambda-1)/\lambda$ which is still a (repelling) fixed point, but the other two are new points of prime period 2.

Are these new periodic points attracting or repelling or nonhyperbolic?

By evaluating the quadratic formula, we obtain the 2 new points of prime period 2:

$$p_0 = \frac{\lambda+1 + \sqrt{(\lambda+1)(\lambda-3)}}{2\lambda} \text{ and } p_1 = \frac{\lambda+1-\sqrt{(\lambda+1)(\lambda-3)}}{2\lambda}$$

when $\lambda > 3$. Also, $\frac{d}{dx}f_\lambda^2(P_0) = f_\lambda'(p_0)f_\lambda'(p_1)$

$$= \lambda(1 - 2p_0) \cdot \lambda(1 - 2p_1)$$

$$= -\lambda^2 + 2\lambda + 4.$$

Thus, p_0 and p_1 are attracting when $3 < \lambda < 1 + \sqrt{6}$, nonhyperbolic for $\lambda = 1 + \sqrt{6}$, and repelling when $\lambda > 1 + \sqrt{6}$.

If the previous pattern from fixed point to 2 points of prime period two repeats, then we would expect each point of prime period 2 to "give birth" to two points of prime period 4 (for a total of 4 points of prime period four) when the period 2 points are nonhyperbolic at $\lambda = 1 + \sqrt{6}$.

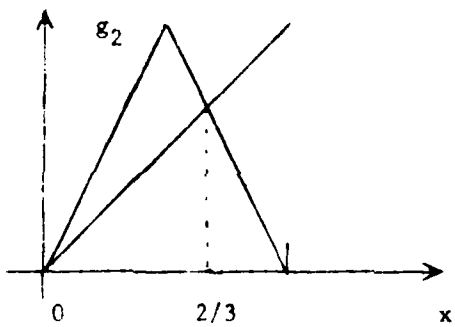
We will now turn to the question: how many periodic points are there?

Example 2. Let $g_2(\lambda) = \begin{cases} 2x, & \text{if } 0 \leq x \leq 1/2 \\ 2(1-x), & \text{if } 1/2 \leq x \leq 1. \end{cases}$

So g_2 is the tent map from the previous section. It is easier to find the periodic points of this map than the logistic map of example 1.

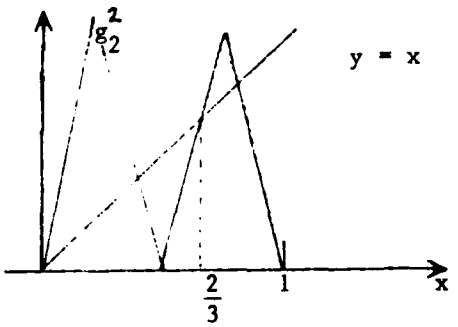
We will develop a geometric technique to find periodic points of different order.

First, periodic points of order 1 are fixed points and are obtained from the intersection of g_2 with the identity map:



So, g_2 has fixed points at 0 and 2/3.

To find points of prime period 2, we graph g_2^2. This can be done without an explicit representation of g_2^2! Note that $g_2^2(x) = g_2(g_2(x))$, so that the range values of g_2 then become the domain values for the next g_2. Thus, since g_2 maps $[0, 1/2]$ onto $[0,1]$, $g_2(g_2(x))$ will go through a complete up and down graph for $x \in [0, 1/2]$ and for $x \in [1/2, 1]$...



Hence, g_2^2 will intersect y = x at 4 points, two of which must be 0 and 2/3. The 2 new points are points of prime period 2.

We can repeat this graphical technique to find periodic points of order n. It is clear that in this case, as $n \to \infty$, we will obtain an infinite number of periodic points. And, the collection of all periodic points will be dense in the interval, [0,1].

One of the properties of g_2 which lends itself to this technique is that the maximum value of g_2 is the same as the maximum value of the domain: one. In example 1 the logistic map, f_λ, will not achieve a maximum value of one until $\lambda = 4$. For this value, f_4 and g_2 have essentially the same behavior. The hard part is tracking their behavior when $\lambda < 4$ (for f_λ) and $\lambda < 2$ for g_λ. (See the exercises.)

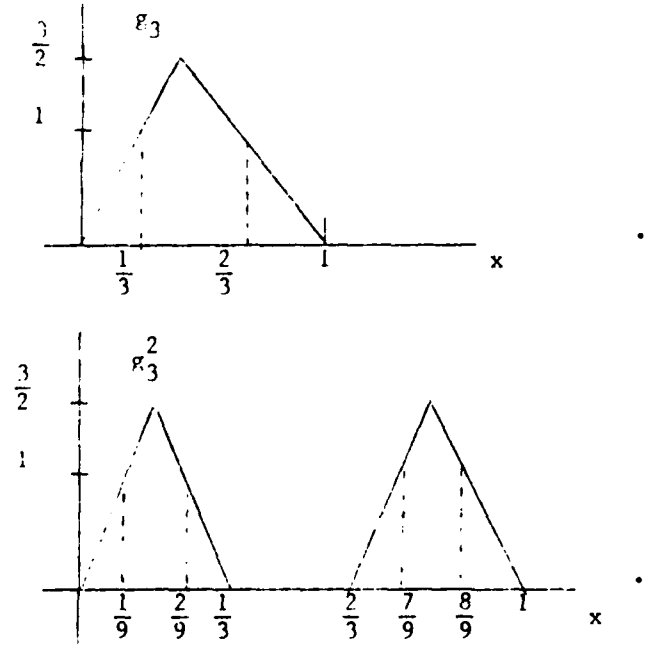
We will now look at the case when $\lambda > 2$ for g_λ, (which is the same as that for f_λ when $\lambda > 4$). In particular, we'll examine $g_3(x)$.

Example 3. Let $g_3(x) = \begin{cases} 3x & , \text{ if } 0 \le x \le 1/2 \\ 3(1-x) & , \text{ if } 1/2 \le x \le 1. \end{cases}$

The most important difference here, from earlier cases, is that when $1/3 < x < 2/3$, $g_3(x) > 1$, and so points between 1/3 and 2/3 get mapped out of the domain interval, [0,1]. Repeated applications of g_3 will take these points to $-\infty$. We could interpret this as a type of resonance. Initial conditions in the

physical system which correspond to these values in the domain of the return map are carried out-of-bounds.

But the values between 1/3 and 2/3 are not the only ones which are attracted to $-\infty$. For example, $x = 1/6$ is mapped by g_3 to 1/2. And now 1/2 will be attracted to $-\infty$. By using the graphical technique in the previous example, we can determine that every point in [0,1] <u>except those in the Cantor Set</u> will eventually be attracted to $-\infty$:



That is, $g_3^n(x)$ will take 2^{n-1} intervals of length $\frac{1}{3^n}$ from its domain, and map them out of [0,1], thus sending them on their way to $-\infty$. So, the points in the Cantor Set are the only points which stay in [0,1] after arbitrary iterations by g_3.

Note that all numbers of the form $\frac{k}{3^n}$ are eventually mapped to zero, and periodic points will be dense in the Cantor Set. (See the exercises.)

Although the number and type of periodic points for g_λ is closely related to the number and type of periodic points for f_δ, (for a suitable choice of δ), there is a fundamental difference. For $\lambda > 1$, g_λ has no attracting periodic points since the slope of g_λ is always $\pm \lambda$. Thus, g_λ is useful since it's easy to work with, but not very realistic, since we expect physical systems to have some steady state solutions.

We'll conclude this section with a remarkable result due to Sarkovskii.

Theorem 3. If f is a continuous function mapping R to R and f has a point of prime period 3, then for every natural number, n, there is a point of prime period n.

This is actually a corollary to Sarkovskii's theorem which gives a precise listing of what periods imply what other periods.

Thus, f_4 and g_2 have points of prime period n for all n. In fact, f_λ has points of prime period 3 for $\lambda = 3.839$.

Exercises

1. What is the degree of the polynomial, $f_\lambda^4(x)$? If $3 < \lambda < 1 + \sqrt{6}$, how many known roots are there to $f_\lambda^4(x) - x = 0$?
2. What is the degree of the polynomial, $f_\lambda^3(x)$? How many known roots are there to $f_\lambda^3(x) - x = 0$?
3. Use numerical methods to find four points of prime period 4 of $f_\lambda(x)$ when $\lambda = 1.01 + \sqrt{6}$.
4. Use numerical methods to find three points of prime period 3 of $f_\lambda(x)$ when $\lambda = 3.839$.

5. Draw a graph of $f_\lambda^4(x)$ when $\lambda = 1 + \sqrt{6}$. (Hint: Use the answer to prob 3 to make the graph more accurate.

6. Prove that $g_3(x)$ maps points of the form $\frac{k}{3^n}$ eventually to zero.

7. Find the two points of prime period 2 of $g_3(x)$. Are these points in the Cantor set? If so, write them in base 3 form using only 0's and 2's.

8. Find the 4 points of prime period four of $g_3(x)$. Are these points in the Cantor set? If so, write them in base 3 using only 0's and 2's.

II.5 Invariant Sets and Attractors

We'll start with a couple of definitions.

Definition 1. A set, F, is an invariant set of the (return) map f if and only if for each $x \in F$, $f(x) \in F$.

Definition 2. A set, P, is an attractor of f if and only if there is an $\epsilon > 0$ so that if p is any point in P and x is any point such that $d(x,p) < \epsilon$ then all limit points of the orbit of x are contained in P.

The definition of invariant set is straightforward. For example, a fixed point is an invariant set, as is any periodic orbit: that is, the collection of m points is invariant. A more interesting example is the Cantor set. In the last section, we saw that the Cantor set is an invariant set of g_3 and $f_{9/2}$. Attractors are fairly easy to understand, but their definition is tedious. Intuitively, a set, P, is an attractor if nearby points are drawn closer and closer by the return map. In the definition, we used "d" to emphasize that distance can be measured in any metric space and we are not limited to return maps whose domain is the real line (or subset thereof). Also, if the orbit of a point approaches ∞, then ∞ is considered to be a limit point of the orbit, and can also be considered an attractor.

Recall from part I that the fixed point of every contraction map is actually an attractor of that map. In fact, an IFS has an attractor which is a fractal, even though we would not usually consider an IFS to be a return map of some physical system.

In a physical system, an invariant set corresponds to trajectories which keep intersecting the Poincaré section in the same set of points. It should be noted that every attractor is also an invariant set, but has the additional physical property of attracting nearby trajectories.

Example 1. Explain the physical significance of the logistic map, f_λ, as λ increases ($\lambda > 1$).

We will give a physical interpretation of the information presented in example 1 of II.4:

For $1 < \lambda < 3$, there is one attracting fixed point at $(\lambda - 1)/\lambda$. All trajectories (except one) are drawn to this fixed point, so the physical system has a stable, attracting 1-cycle as its steady state solution. Note that zero is a repelling fixed point, so it can be considered as an unstable invariant set. The one trajectory which intersects the Poincaré section at zero is the only one which isn't attracted to $(\lambda - 1)/\lambda$. For this reason, zero is ignored.

For $3 \leq \lambda < 1 + \sqrt{6}$, there is one attracting set which is an orbit of period two. The fixed point at $(\lambda - 1)/\lambda$ is now a repellor, and hence can be ignored. For all practical purposes, the physical system will draw all trajectories into a stable steady state solution of a 2-cycle. This 2-cycle was generated through a bifurcation of the fixed point $(\lambda - 1)/\lambda$.

As λ increases through $1 + \sqrt{6}$, the stable 2-cycle above will bifurcate into a stable 4-cycle which becomes the steady state solution of the physical system. Now we are ignoring two repelling fixed points--at 0 and $(\lambda - 1)/\lambda$ --and two repelling points of prime period 2. For fairly obvious reasons, including noise, it is impossible for an actual trajectory in the system to repeat an unstable or repelling cycle (or orbit).

This bifurcation behavior will continue for smaller and smaller increases in the parameter, λ. The accumulation point will correspond to a type of chaos. At this chaotic value of λ there will be no attracting cycle, so the motion in the system will be irregular and aperiodic. There can still be a steady state solution (attractor), but it will not be periodic.

Further increases in λ will show periods of chaos interspersed with regular, attracting cycles. When the cycles exist, they will bifurcate as before and lead to another period of chaos. When $\lambda = 4$, λ will be so large that no attracting solution will exist, even an aperiodic one. At this value, there is no attractor and the system is completely turbulent.

Exercises.

1. For $\lambda > 2$, find the measure of the invariant set of g_λ. [Hint: Add up the length of the intervals in the complement of the invariant set, then take one minus this.]
2. Can g_λ ($\lambda > 1$) be used to model a physical system? Explain.

II.6 Scrambled Sets--Definitions of Chaos

We will give two definitions of chaos in this section. To do so, the concepts of "sensitive dependence on initial conditions" and "scrambled sets" need to be clarified.

Definition 1. A function (return map) has sensitive dependence on initial conditions if and only if there is an $\varepsilon > 0$ so that for every x and every $\delta > 0$ there is a y and n such that $d(x,y) < \delta$ and $d\left(f^n(x), f^n(y)\right) > \varepsilon$.

The definition states that arbitrary accuracy of orbits cannot be maintained. That is, there is an $\varepsilon > 0$ so that no matter how close you start to x, there is a y which will end up ε units away from x after some number of iterations.

This behavior is strengthened (worsened?) by scrambled sets:

Definition 2. Suppose $f: I \to J$ is a continuous function mapping I onto J, where J is a subset of I. Let λ be the length of J. S is a scrambled set of f if and only if for every x and y in S, $(x \neq y)$,

(i) $\limsup_{n \to \infty} |f^n(x) - f^n(y)| = \lambda$, and

(ii) $\liminf_{n \to \infty} |f^n(x) - f^n(y)| = 0$.

We'll explain "lim sup" and "lim inf": the lim sup as $n \to \infty$ is the supremum of values achieved by $|f^n(x) - f^n(y)|$ as $n \to \infty$. Thus, condition (i) implies that any two points in S will be iterated as far apart as possible (J is the limiting factor). At the same time, condition (ii) implies that these two points will be iterated close together again as well. The terminology, "scrambled", seems understated.

First, note that sensitive dependence on initial conditions happens for every value in the domain, I. (We used arbitrary metric space notation in definition 1 so that it can be applied to higher dimensional situations.) But the behavior in a scrambled set is localized to that set, and it will never be the entire domain. (A metric, d, could also be used in definition 2.)

Second, note that if a return map has an attracting set, it cannot have sensitive dependence on initial conditions since points drawn to the attractor will stay close together.

We need one more definition.

Definition 3. A (return) map, f, is transitive (or nomadic) if and only if there is an x so that the orbit of x is dense in the domain of f.

We have seen this behavior for g_2 back in section II.3.

Now we'll present two definitions of chaos:

Definition 4. A function, f, is chaotic-3 if and only if the following 3 conditions are satisfied:

(i) f has sensitive dependence on intial conditions,

(ii) f is transitive, and

(iii) the periodic points of f are dense in the domain.

Definition 5. A function, f, is chaotic-s if and only if there is a scrambled set of uncountable cardinality.

In the article "On Scrambled Sets for Chaotic Functions", Andrew Bruckner and Thakyin Hu showed that if we assume the continuum hypothesis (see Fundamentals of Contemporary Set Theory, by Devlin) then a function, f, is chaotic-s if and only if the second iterate, f^2, is transitive. They also showed that g_2 (the tent map) is chaotic-s.

Example 1. The tent map, g_2, is also chaotic-3.

We only need to show that g_2 has sensitive dependence on initial conditions. For this function, we can choose $\varepsilon > 0$ to be any number less than 1. Fix such an ε. Now let $\delta > 0$ and choose m so that $\frac{1}{2^m} < \delta$. Fix a number, x, in $[0,1]$ and write x in base 2 as $x = .d_1 d_2 \ldots$ where each $d_i \in \{0,1\}$. Now, we need to find a y within δ of x so that after some number of iterations $f^n(x)$ and $f^n(y)$ will be at least ε apart:

Define $y \in [0,1]$ as $y = .e_1 e_2 \ldots$ where $e_i = d_i$ for $1 \leq i \leq m$ and $e_i \neq d_i$ for all $i > m$. Thus, $y - x$ has zeros in the first m positions; and hence, $|y - x| < \frac{1}{2^m} < \delta$.

Now, $g_2(z) = g_2(.z_1 z_2 \ldots) = \begin{cases} .z_2 z_3 \ldots & \text{if } z_1 = 0 \\ .q_2 q_3 \ldots & \text{if } z_1 = 1 \end{cases}$ where $q_i \neq z_i$ for each i.

Thus, after iterating x and y m times, $g_2^m(x) = .x_{m+1} x_{m+2} \ldots$ where either each $x_{m+j} = d_{m+j}$ or each $x_{m+j} \neq d_{m+j}$, and similarly for y: $g_2^m(y) = .y_{m+1}\, y_{m+2} \ldots$ where each $y_{m+i} \neq x_{m+i}$. This last fact implies that $|g_2^m(y) - g_2^m(y)| = .\bar{1}$ base 2 $= 1$. So, since $\varepsilon < 1$, we have shown that g_2 has sensitive dependence on initial conditions.

In section II.3, it was shown that g_2 is transitive and that the periodic points of g_2 are dense in $[0,1]$. Thus, g_2 is chaotic-3.

Example 2. The logistic map, f_λ, is chaotic-3 on its invariant set when $\lambda \geq 2 + \sqrt{5}$.

First, note that we are restricting the domain of f_λ to its invariant set since if $\lambda > 4$ (as is $2 + \sqrt{5}$) then f_λ will take the majority of the interval $[0,1]$ off to $-\infty$. Recall that the invariant set will be a Cantor-type set (the actual Cantor set at $\lambda = 9/2$) when $\lambda > 4$.

Using symbolic dynamics, it is possible to show that f_λ has a dense orbit (in its invariant set) and that periodic points are dense (in its invariant set), whenever $\lambda > 4$.

Now, when $\lambda \geq 2 + \sqrt{5}$, $|f'_\lambda(x)| \geq 1$ for all x in the invariant set. (See the exercises.) Thus, there can be no attracting set for f_λ, and hence f_λ must have sensitive dependence on initial conditons. So f_λ is chaotic-3 when $\lambda \geq 2 + \sqrt{5}$. In fact, f_4 is also chaotic-3. (f_4 and g_2 are "topologically conjugate." See II.10.)

We will now discuss the physical implications of our criteria for chaos. Sensitive dependence on initial conditions is quite plausible physically since it precludes an attracting set. It is harder to rationalize the necessity of dense periodic points. But, since the orbits cannot be attracting, knowing that there are unstable m-cycle trajectories intersecting the Poincaré section doesn't hurt. The transitivity says that there is a trajectory which will intersect the Poincaré section in every interval, which is certainly a type of irregularity.

If there is an uncountable scrambled set, that shows that many trajectories are repeatedly converging and diverging along the Poincaré section. That this also implies transitivity lends physical credibility to the chaos-s definition.

In general, it is not easy to determine if a return map is either chaotic-3 or chaotic-s. One usually uses numerical techniques to see if it <u>might</u> be chaotic and then conjectures one way or another. See section II.10.

<u>Exercises</u>.

1. Prove that if $\lambda \geq 2 + \sqrt{5}$ then $|f'_\lambda(x)| \geq 1$ on its invariant set.
2. Prove that if f is any function with sensitive dependence on initial conditions, then f cannot have an attracting set.
3. Do you think that f_λ, the logistic map, can be chaotic for values of $\lambda < 2 + \sqrt{5}$? Where is it definitely <u>not</u> chaotic?

II.7 Universality--The Bifurcation Diagram

First, we'll describe how to interpret a bifurcation diagram. (See page 96.) The non-linearity parameter, λ for the logistic map: $f_\lambda(x) = \lambda x(1-x)$, is plotted along the horizontal axis. The attracting set for that particular λ is plotted along the vertical axis, which runs from zero to one for f_λ.

Thus, for $1 < \lambda < 3$, $(\lambda - 1)/\lambda$ is the attracting set (1-cycle). At $\lambda = 3$, this bifurcates to give an attracting 2-cycle, etc.

Where whole intervals seem to be shaded along the vertical axis, there is no attracting cycle, but rather attracting intervals (or subsets thereof). Notice that there are bands of attracting cycles (periodicity) interspersed among the aperiodic regions. When λ is equal to 4, the whole interval, $[0,1]$, is shaded and chaos-3 is in effect. The shaded areas prior to this do not correspond to chaos-3 (since there will not be a dense orbit) but do correspond to a weak chaos or aperiodic behavior. (Weak turbulence in fluid dynamics.)

We will present universality from the standpoint of the bifurcation diagrams. The figures in this section may be helpful. (Pages 98 to 101.)

Our first discussion will center on the constant, $\delta = 4.6692016\ldots$. If we denote by L_i the values of λ for which the logistic map, $f_\lambda(x) = \lambda x(1 - x)$, has an attracting 2^i-cycle, then L_i will be an interval.

For example, f_λ, has an attracting 2-cycle for $1 < \lambda < 3$, so $L_0 = (1,3)$. Similarly, $L_1 = (3, 1 + \sqrt{6})$. Define Δ_i to be the length of L_i. Then, Mitchell Feigenbaum has shown that $\lim_{i \to \infty} \frac{\Delta_i}{\Delta_{i+1}} = \delta$.

The universal aspect of δ is that Feigenbaum's formula will hold for any system with a bifurcation diagram. Essentially, every bifurcation diagram looks the same when rescaled along the horizontal axis. Thus, the cascading of 2^n-cycles (or any bifurcating cycles) will be the same for the logistic map and for the experiment with a heated fluid.

In fact, the windows in the bifurcation diagrams which are rescaled and blown up in this section, can be seen to be essentially the same as well. All bifurcating cycles do so at the same rate (in the limit).

We will now discuss the constant, $\alpha = 2.50290787...$. Notice that on the bifurcation diagram a one-cycle literally splits into a two-cycle, then each branch of the two-cycle splits again to give a total of a four-cycle, etc. Denote the vertical distance between the two branches of the 2-cycle at their point of bifurcation as A_1. Pick a pair of the new split branches in the 4-cycle. (It doesn't matter whether we choose the upper or lower pair. But the branches of the pair must have originated from the same branch of the 2-cycle.) For this pair of branches, denote by A_2 the vertical separation when each one bifurcates (to give an 8-cycle). If we continue to find A_i's in this fashion, then Mitchel Feigenbaum also showed that $\lim_{i \to \infty} \frac{A_i}{A_{i+1}} = \alpha$. This fact is independent of which branches one chooses, it only depends on the branches bifurcating.

The universal aspect of α is that all bifurcation diagrams, (for any system), are now essentially the same when rescaled along the vertical axis as well! The combination of α and δ show that every bifurcation diagram has essentially the same rates of vertical and horizontal accumulation.

At this point, we will describe the technique for generating a bifurcation diagram for $f_\lambda(x) = \lambda x(1 - x)$...

Essentially, we would like to graph the attracting cycle (if it exists) for different values of λ. Recall that the cycle is attracting if $\left| \frac{d}{dx} f^n(x_0) \right| < 1$, and that $\frac{d}{dx} f^n(x_0) = \prod_{i=0}^{n-1} f'(x_i)$ where $\{x_0, x_1, \ldots, x_{n-1}\}$ is the n-cycle. Thus, the cycle is attracting any time $\frac{1}{2}$ is one of the elements of the n-cycle since $f'(\frac{1}{2}) = 0$, implying that $\left|\frac{d}{dx} f^n(x_0)\right| = 0 < 1$.

In the evolution of an n-cycle, (being created from an $\frac{n}{2}$-cycle, and becoming a 2n-cycle), the n-cycle will go from weakly attracting-- $\frac{d}{dx} f_\lambda^n(x_0) = 1$ --to stable, to weakly attracting again. Essentially, $\frac{d}{dx} f_\lambda^n$ will achieve values from 1 down through 0, and then to -1 when it bifurcates. Thus, $\frac{1}{2}$ will be an element of every stable n-cycle in the bifurcation diagram.

For this reason, we use $\frac{1}{2}$ as the starting point, (for a fixed λ), and iterate some number of times, say up to $f^{50}(\frac{1}{2})$. Now, if there is an attracting n-cycle for the value of λ with which we're working, then $f^{50}(\frac{1}{2})$ should be "attracted" to it. Thus, we plot on the graph $f^{50+i}(\frac{1}{2})$ for i equal from one to 50. These fifty plotted points will be on the attracting n-cycle if it is present, or, they will bounce around in some attracting set. We plotted several thousand points in the diagrams in this section in order to completely (or partially) shade the aperiodic attracting sets.

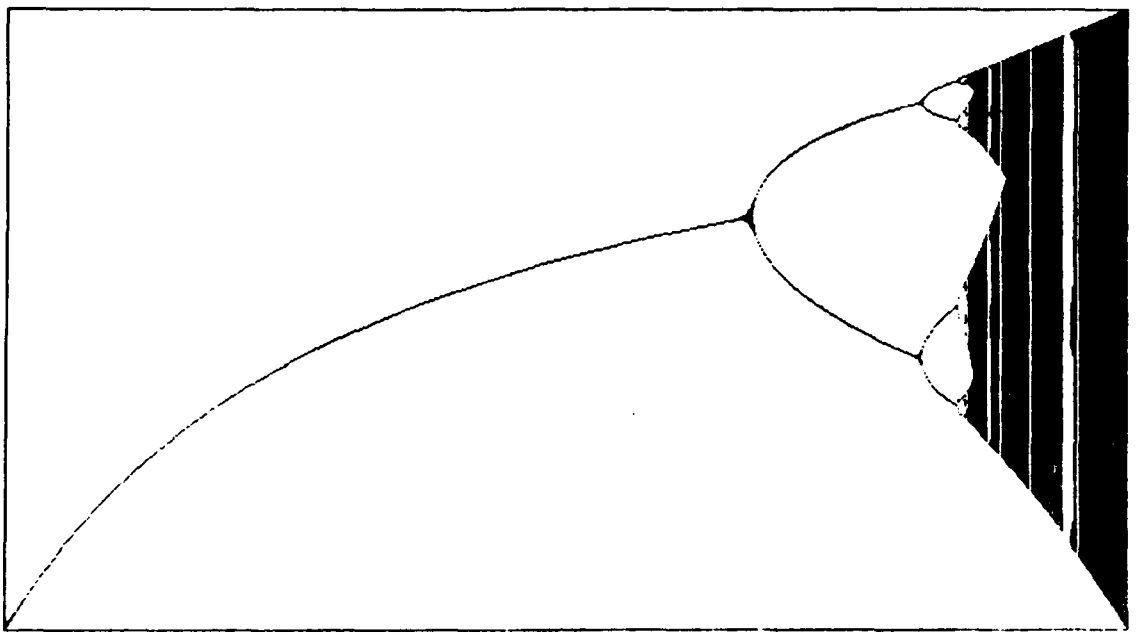


Figure 16

"Figure 16 is the bifurcation diagram of the logistic map $f_\lambda(x) = \lambda x(1 - x)$. Values of λ between 1 and 4 are plotted along the horizontal axis and the attracting set is plotted along the vertical scale which is from 0 to 1.

The largest "window" in the predominantly shaded region to the right is rescaled in figure 17 to show a bifurcating 3-cycle.

This same figure is also on page 67."

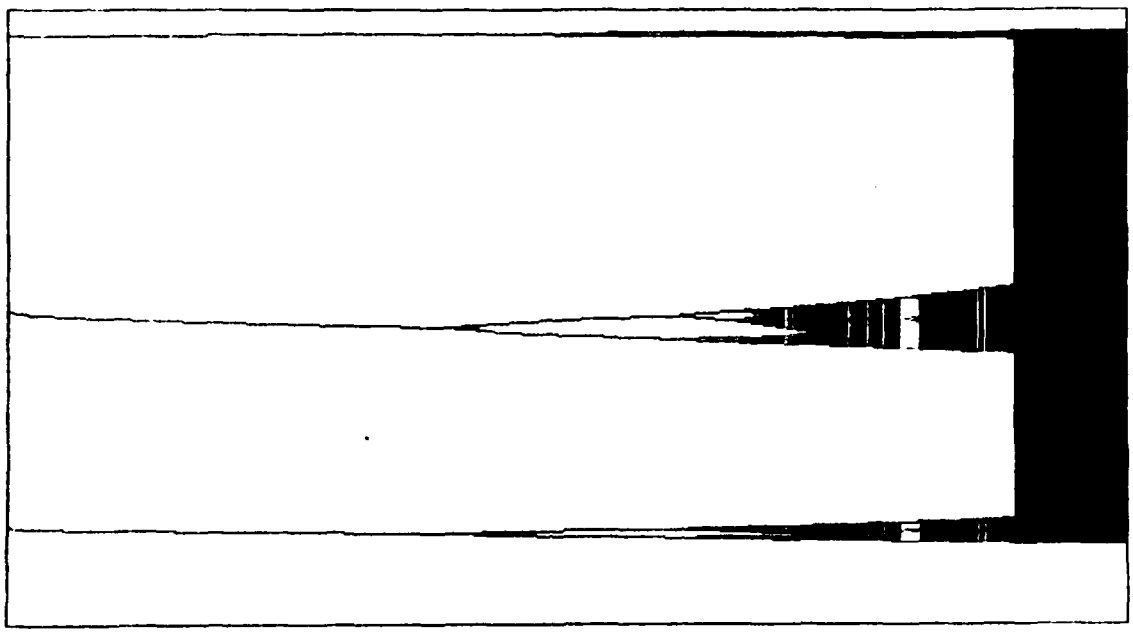


Figure 17

"Figure 17 shows a 3-cycle bifurcating into a 6-cycle, then 12-cycle, etc.

The bifurcations are simultaneous even though it looks as if the upper branch takes longer to bifurcate. This is because of the resolution of the graphics.

The vertical scale is still 0 to 1. But, λ is now between 3.8284 and 3.86 along the horizontal axis.

The largest "window" is rescaled in figure 18. It is a 9-cycle."

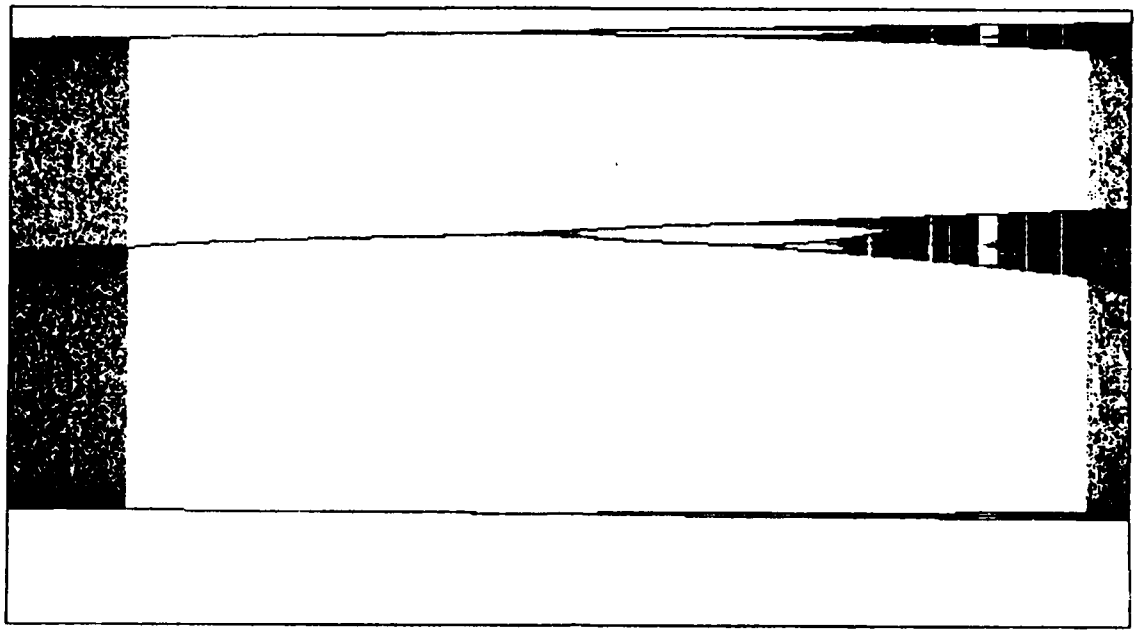


Figure 18

"Figure 18 is a rescaling of the largest window in figure 17. The vertical axis is also rescaled so that only the large middle window is visible, and hence this looks like a 3-cycle, but is just $\frac{1}{3}$ of a 9-cycle.

λ is between 3.85355 and 3.85415, and the vertical scale is between 0.4324 and 0.5405.

The largest middle window is again rescaled in figure 19."

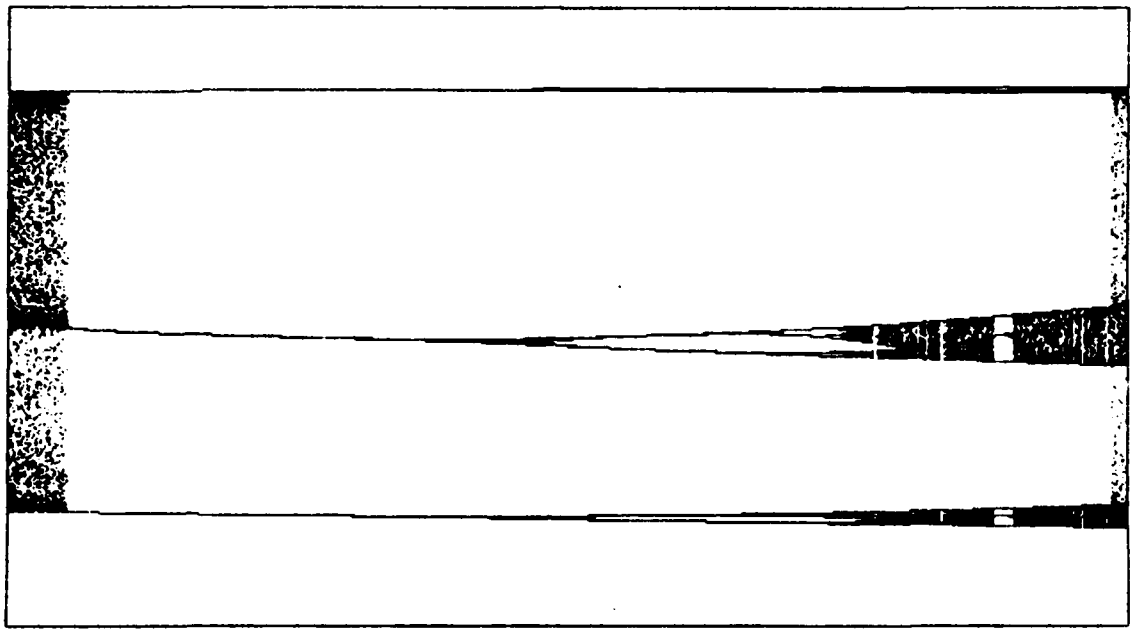


Figure 19

"Figure 19 is $\frac{1}{9}$th of a 27-cycle. The values of λ are between 3.854069 and 3.854079. The vertical axis is scaled between 0.493754 and 0.506902.

Notice the striking similarity between figures 17, 18, and 19. Figure 18 looks like the vertical mirror image of figure 17, and figure 19 is the vertical mirror image of figure 18."

In section II.6 we presented a theorem due to Sarkovskii. Here we will give Sarkovskii's ordering of the natural numbers.

Definition. The Sarkovskii ordering of the natural numbers is:

$$3 < 5 < 7 < 9 \ldots < 2\cdot 3 < 2\cdot 5 < 2\cdot 7 \ldots$$

$$< 2^2\cdot 3 < 2^2\cdot 5 < 2^2\cdot 7 < \ldots < 2^3\cdot 3 < 2^3\cdot 5 < \ldots$$

$$< \ldots < 2^3 < 2^2 < 2 < 1.$$

Thus, one first lists all odd numbers, in the usual order, then all products of an odd number and 2, then an odd number and 2^2, etc. This will list all the natural numbers except those that are powers of 2 (and 1). List these powers of 2 last, in reverse order.

A more general Sarkovskii's theorem is:

Theorem. Suppose $f: R \to R$ is continuous. If f has a point of prime period k and $k < m$ in Sarkovskii's ordering, then f also has a point of prime period m.

Thus, if f has a 2-cycle, f will also have a 1-cycle and if f has a 3-cycle, f will have an n-cycle for every value of n. However, this theorem doesn't say the cycles are attracting.

For example, when $\lambda = 3.1$, f_λ has a stable 2-cycle, so it must also have a 1-cycle. Well, $(\lambda-1)/\lambda$ is still a fixed point when $\lambda = 3.1$, but it is repelling.

So when $\lambda = 3.83$ there will be a stable 3-cycle, and hence an n-cycle for every n. But for this λ, only the 3-cycle is attracting; every other n-cycle is repelling, and thus, not drawn on the bifurcation diagram.

Exercises.

1. For each λ between 1 and 4, find the range of f_λ.

2. For each λ between 1 and 4, define $g_0(\lambda) = \frac{1}{2}$, $g_1(\lambda) = f_\lambda(\frac{1}{2}) = \frac{\lambda}{4}$, $g_2(\lambda) = f_\lambda(g_1(\lambda))$, etc. Show that $g_0(2) = g_1(2) = \ldots = \frac{1}{2}$.

3. Show that whenever $\frac{1}{2}$ has period n, then $g_1(\lambda)$, (as in prob 2), is tangent to the bifurcation diagram at that λ.

4. Show that whenever $\frac{1}{2}$ has period n, then $g_2(\lambda)$, (as in prob. 2), is tangent to the bifuraction diagram at that value of λ.

5. For $\lambda \geq 2$, show that if $0 < x < g_2(\lambda)$, then $\lim_{n\to\infty} f_\lambda^n(x) \geq g_2(\lambda)$ and, for all x, $\lim_{n\to\infty} f_\lambda^n(x) \leq g_1(\lambda)$, where g_1 and g_2 are as in problem 2.

6. If g_i is as in problem 2, what can be said about each $g_i(\lambda)$ for those λ where $\frac{1}{2}$ is an element of some n-cycle?

II.8 Higher Dimensions

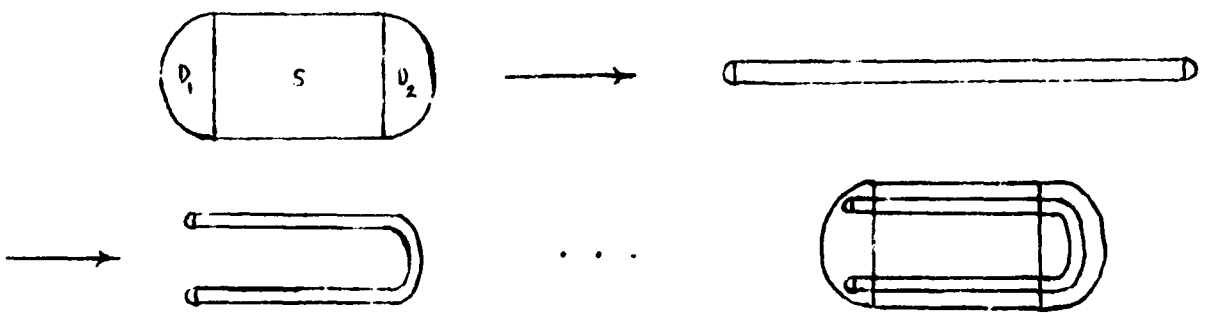
When we use a one-dimensional Poincaré section there are two types of bifurcations which the return map can display. The one we looked at in section II.8 was a period-doubling bifurcation, and this occurs when $\frac{df^n}{dx}(x_0) = -1$, (where x_0 is any point in an n-cycle).

When $\frac{df^n}{dx}(x_0) = 1$, the return map goes through a saddle-node bifurcation. Essentially, a saddle-node attracts points from one side while repelling points on the other side. For example, $h_\lambda(x) = e^x - \lambda$ goes through a saddle-node bifurcation as λ increases through 1.

In a phase space of dimension greater than two, we will probably be using a Poincaré section of dimension larger than one. Thus, the return map is not a function of one variable, but rather of several variables. The types of bifurcations that can occur (both in the phase-space and on the Poincaré section) are varied. They include analogies of the period-doubling and saddle-node bifurcations as well as others, the most common of which is the "Hopf bifurcation".

As an example of the type of dynamics possible, we'll give a brief treatment of the Horseshoe map, due to Smale.

Example 1. The Horseshoe map takes the figure drawn into itself by stretching, contracting and bending:



This map works on a 2-dimensional region in the plane and maps it into itself. We've defined it geometrically because it is easier to work with that way.

Since the Horseshoe map shrinks (and stretches) the set D_1 into a subset of D_1, it is a contraction mapping, and thus has a unique fixed point in D_1. Also, all points in D_2 are mapped into a subset of D_1; so, for every point $x \in D_1 \cup D_2$, $\lim_{n\to\infty} f^n(x) = x_0$ where x_0 is the unique fixed point of f, the Horseshoe map, in D_1.

Many of the points in S will also be mapped into D_1 and thus iterate towards x_0 as well. But there is a two-dimensional Cantor-type set in S which is invariant. That is, the Horseshoe map leaves this 2-dimensional Cantor-type set fixed.

Points in this invariant set can be shuffled around by the map, but will always remain inside it. This is exactly analogous to the logistic map, f_λ, when $\lambda > 4$. There, we had a one-dimensional set which was invariant. (Also for the tent map, g_λ, when $\lambda > 2$.)

Next we'll look at a function algebraically called the Henon map.

Example 2. Define $H_{a,b}: R^2 \to R^2$ by: $H_{a,b}(x,y) = (a - by - x^2, x)$.

Thus, points in R^2 are mapped to points in R^2. There is an analogy to hyperbolicity of one-dimensional functions which relies on full derivatives of multi-dimensional functions:

$$DH_{a,b}(x,y) = \begin{bmatrix} \frac{\partial H^1}{\partial x}(x,y) & \frac{\partial H^1}{\partial y}(x,y) \\ \frac{\partial H^2}{\partial x}(x,y) & \frac{\partial H^2}{\partial y}(x,y) \end{bmatrix}$$

where $H^1(x,y) = a - by - x^2$ and $H^2(x,y) = x$.

Thus, $DH(x,y) = \begin{bmatrix} -2x & -b \\ 1 & 0 \end{bmatrix}$.

The Jacobian of H is simply the determinant of the derivative of H, $|DH| = b$.

H is hyperbolic so long as $|b| \neq 1$. When $b = 0$, H is no longer dependent on y, and is essentially: $H_a(x) = a - x^2$, which is analogous to the logistic map.

When $b \neq 0$, we can algebraically find the inverse of H to obtain:

$H^{-1}_{a,b}(x,y) = \left(y, \frac{a - x - y^2}{b}\right)$ which looks quite similar to H itself.

We will try to find the fixed points of $H_{a,b}$, assuming that $0 < |b| < 1$:

First, set

$$(x,y) = (a - by - x^2, x),$$

which implies $x = y$ and $x = a - bx - x^2$, or $x^2 + (1 + b)x - a = 0$. Thus, this has real solutions if and only if $(1 + b)^2 + 4a \geq 0$.

So, when $a < \frac{-1}{4}(1+b)^2$, there are no fixed points; when $a = \frac{-1}{4}(1 + b)^2$, there is one fixed point at $x = \frac{-1}{2}(1 + b)$; and when $a > \frac{-1}{4}(1 + b)^2$, there are two fixed points at: $x = \frac{-(1 + b) + \sqrt{(1 + b)^2 + 4a}}{2}$ and $x = \frac{-(1 + b) - \sqrt{(1 + b)^2 + 4a}}{2}$. Note that the y-coordinates of all these points are the same as the x-coordinates by our substitutions.

The critical value of a, a_{crit}, is $\frac{-1}{4}(1 + b)^2$. On the line $y = x$, we have has no fixed points of $H_{a,b}$ when $a < a_{crit}$, then 1 fixed point for $a = a_{crit}$, and then 2 fixed points for $a > a_{crit}$. It turns out that of these two fixed points, one is a saddle point and the other is (sometimes) attracting. Thus, as a increases through a_{crit}, a saddle-node bifurcation occurs. As a increases further, one of the fixed points bifurcates into a period-2 point. This is a period-doubling bifurcation.

The dynamics of the Henon map get very complicated. See An Introduction to Chaotic Dynamical Sytems for further details.

"Strange Attractors" are quite common for higher dimensional return maps. When a = 1.4 and b = -0.3, the Henon map, $H_{a,b}$, seems to have one. Strange attractors are attracting sets for a return map which display fractal-type properties. There are strange attractors imbedded in 3-dimensional toroidal figures for some systems, and the bifurcation diagram itself is a type of strange attractor. This is a nice relationship between fractals and chaos: dynamical systems can generate fractal sets. For example, an IFS is a type of dynamic system. Also, fractal sets display "chaotic organization" to the average observer.

Higher dimensional dynamics is still a frontier for researchers and there are many unsolved problems: both specific ones and general ones.

II.9 Complex Dynamics

Because of the elegant theory of complex numbers, dynamics which occur in R^2 are often interpreted to lie in the complex plane. We'll spend most of this section discussing the Mandelbrot set, which is a parameter space map for a simple dynamical system in complex variables.

First, recall that a complex number can be represented as: $a + ib$ or $re^{i\theta}$, where a, b, r, and θ are real and $i^2 = -1$. When graphing complex numbers the number, a, goes on the x-axis (the real axis) and the number, b, goes along the y-axis (imaginary axis). Plotting $a + ib$ is the same as plotting (a,b). The form, $re^{i\theta}$, is analogous to polar coordinates since $e^{i\theta} = \cos\theta + i \sin\theta$. That is, to graph $re^{i\theta}$, just graph (r,θ) in the polar plane.

To mulitply: $(a + ib)(c + id) = ac - bd + i(ad + bc)$, and $(re^{i\theta})(qe^{i\theta}) = rqe^{i(\theta+\phi)}$. A multiplicative inverse of $a + ib$ is $\frac{1}{a + ib} = \frac{a - ib}{a^2 + b^2}$, since $(a + ib) \cdot \left(\frac{a - ib}{a^2 + b^2}\right) = \frac{a^2 + b^2}{a^2 + b^2} = 1$, (assuming $a^2 + b^2 \neq 0$).

The modulus of a complex number, $z = a + ib$, is $|z| = |a + ib| = \sqrt{a^2 + b^2}$. If $z = re^{i\theta}$, then $|z| = |r|$, which is the distance from the complex number, z, to the origin, $0 = 0 + 0i$.

Before we present the algorithm for the Mandelbrot set, we'll look at its real number analog.

Example 1. Define $h_c(x) = x^2 + c$. Show that h_c has essentially the same dynamics (as a return map) as the logistic map, f_λ, for appropriate values of λ and c.

First, we'll choose $1 < \lambda < 4$ and let $c = \frac{\lambda}{4}(2 - \lambda)$, so that $\lambda = 1 + \sqrt{1 - 4c}$. Now define the linear function, $g(x) = ax + b$, where $a = \frac{-1}{\lambda}$ and $b = \frac{1}{2}$. Then $g \circ h_c = f_\lambda \circ g$, or $g \circ h_c \circ g^{-1} = f_\lambda$. This relationship between h_c and f_λ is called topological conjugacy. It is analogous to two matrices being similar (and thus having the same eigenvalues). It is easy to verify that h_c duplicates the dynamics of f_λ(as λ goes from one to four) as c goes from $\frac{1}{4}$ to -2. (See the exercises.)

One way to check this is to notice that: $(g \circ h_c \circ g^{-1})^2 = g \circ h_c^2 \circ g^{-1}$. Thus, $f_\lambda^n = g \circ h_c^n \circ g^{-1}$ and $g^{-1} \circ f_c^n \circ g = h_c^n$. So, if x_0 is a point of period n for f_λ, then $g^{-1}(x_0)$ is a point of period n for h_c. Also, $\frac{d}{dx} h_c^n(g^{-1}(x_0)) = \frac{d}{dx} f_\lambda^n(x_0)$, which shows that the dynamics of the two functions are the same for all attracting cycles.

If we were to make a bifurcation diagram of h_c, then we would choose zero as the point to iterate, since it will be present as a periodic point in all stable n-cycles. (Just as $\frac{1}{2}$ was present in all stable n-cycles for f_λ.)

Therefore, h_c (which goes through a saddle-node bifurcation for $c = \frac{1}{4}$ (just like f_λ does when $\lambda = 1$) has an attracting fixed point for c between $\frac{1}{4}$ and $\frac{-3}{4}$, an attracting 2-cycle for c between $\frac{-3}{4}$ and $\frac{-5}{4}$, etc; with chaos-3 at $c = -2$.

We'll now develop the algorithm to construct the Mandelbrot set.

Example 2. Define $H_c(z) = z^2 + c$ to be a complex-valued function of the complex variable, z. The nonlinearity parameter, c, is also complex valued.

For each value of c, we determine whether the orbit of zero (a critical point as in example 1) will converge to infinity.

If the orbit does not go to ∞, then c is defined to be in the Mandelbrot set, M. So

$$M = \{c \in C: \lim_{n\to\infty} H_c^n(0) \neq \infty\},$$

where C is the set of all complex numbers.

It is easy to check that zero is in M, since if $c = 0$, then $H_0^n(0) = 0$ for every n. Similarly, one is not in the Mandelbrot set, since, $H_1(0) = 1$, $H_1^2(0) = 2$, $H_1^3(0) = 5$, ... which converges to ∞.

It turns out that if there is an n so that $|H_c^n(0)| > 2$, then c will not be in the Mandelbrot set.

On the real axis between $\frac{1}{4}$ and -2, H_c has the same dynamics as h_c in example 1. Thus, the real numbers from $\frac{1}{4}$ to -2 are all contained in the Mandlebrot set.

The points not in the Mandlebrot set form what is known as a Julia set. The boundary of the Julia set is the same as the boundary of the Mandelbrot set. The Mandelbrot set is a map in parameter space. It describes what parameters give zero a bounded orbit under F_c. The dynamics of F_c in the complex plane (not parameter space) are completely different. There are many possibilities for coming up with interesting pictures.

Other examples can exhibit multiple strange attractors, their corresponding basins, and an invariant set all for the same complex function! By using different coloring schemes, the computer graphics generated by these functions can be amazing and appear frequently on book covers and in magazine articles.

Exercises

1. Verify the second paragraph in example 1.

2. By writing $z = x + iy$, write $H_c(z)$ as a function of x,y ... $H_c(x,y) = (H^1, H^2)$, where F^1 and F^2 are real-valued functions of x and y. (Let $c = c_1 + ic_2$.)

3. Investigate the dynamics of the real-valued function $f_c(x) = \frac{4}{3}x^3 + cx$ for $c \in R$.

4. Using problem 3, what can you say about the complex dynamics of $F_c(z) = \frac{4}{3}z^3 + cz$.

II.10 Randomness

We have often referred to chaotic behavior as being unpredictable. A chaotic physical system is one which is going through seemingly irregular motion. Even though we have very precise definitions of chaotic functions (chaotic-3 and chaotic-s), they are based on characteristics which imply a kind of randomness.

We will now quote some references on the definition of random:

1. From: An Introduction to Information Theory, Pierce. "Random: Unpredictable."
2. From: An Encyclopedic Dictionary of Mathematics. "Practically, random numbers are those that are generated by complex finite algorithms that produce a finite sequence of numbers that have no apparent regularities and are not rejected by tests of typical statistical hypotheses on independence, uniformity, and goodness of fit."
3. From: Mathematics Dictionary, 4th ed., James/James. "Random Sequence: A sequence that is irregular, nonrepetive and haphazard.... A completely satisfactory definition of random sequence is yet to be discovered."
4. A paraphrase of Andrei Kolomogorov (who laid the foundations for modern probability theory in 1933): A finite sequence is random if the shortest algorithm which can generate it is of the same approximate length as the finite sequence.

Mark Kac was a famous mathematician who used to give very popular talks on "randomness." His thesis was that there is no statistical definition of random-

ness. He claimed, (correctly), that given any statistical test--such as those mentioned at the end of the second quote--there is an algorithm which will generate pseudorandom numbers that pass that statistical test.

In other words, if you know that a statistician is going to run some analysis techniques on data, you can give him two sets of data: one set is generated by what we commonly think of as a random process--such as a scatter pattern of arrows on a target--and the other set is generated by a deterministic algorithm. The statistician will not be able to tell if either set of data is "truly random".

It should be noted that although probablity and statistics seem inextricably connected to randomness, the foundations of both subjects rely on "random variables", which are essentially just normal everyday functions. A "random sample" is usually meant to connote a theoretical "random sequence" and is defined in such a way as to make the theory progress smoothly. However, the statistical tests which try to verify this type of randomness are inconclusive. This does not lessen the utility of probability or statistics--they have proven themselves time and again in such diverse areas as gambling and quantum mechanics--it merely points out that "random" might be a term so basic, that it defies definition. This thought is echoed in the third definition.

Finally, we need to peruse Kolomogrov's definition. A full treatment would involve some theoretical computer science, so we'll stick to practicalities. It is theoretically possible to find the algorithm which generates psuedorandom numbers, but it is impossible in practice. Actually, it is just as likely that an algorithm will be found to generate a "truly random" sequence.

Let's consider a chaotic map as a pseudorandom number generator. The two simplest examples are $g(x) = \begin{cases} 2x, & \text{if } 0 \leq x \leq \frac{1}{2} \\ 2(1-x), & \text{if } \frac{1}{2} \leq x \leq 1 \end{cases}$ and $f(x) = 4x(1-x)$: the tent map and logistic map we've already dealt with. As discussed earlier, any finite decimal will be iterated to zero in a finite number of steps (approximately the length of the decimal expansion) by the tent map. So $g(x)$ does not seem to be suitable for generating long strings of pseudorandom numbers.

But $f(x)$, the logistic map, will iterate a point in the interval $[0,1]$ in a seemingly random fashion with few exceptions. (Zero, one-half, and one, as well as the inverses of one-half will all iterate to zero.) Of course, if we just pick a point, say $\frac{1}{\pi}$, and print its orbit, it is obviously not random. If we denote $\{f^n(\frac{1}{\pi})\}_{n=0}^{N}$ by $\{x_0, x_1, x_2, \ldots, x_N\}$ where $x_0 = \frac{1}{\pi}$, $x_1 = f(\frac{1}{\pi})$, etc., then a plot of the points (x_i, x_{i+1}) in the xy-plane will give good graphical evidence that this sequence is highly correlated. This is obvious: each $x_i = f(x_{i-1})$, the graph will fill in points on the curve of $y = f(x)$.

Being a bit more clever, we could choose a number, n, and let each $x_i = f^n(x_{i-1})$. For a value of n greater than the number of significant digits carried by one's calculator or computer, a graphical plot of (x_i, x_{i+1}) will no longer fit on an obvious graph, and thus roundoff error will destroy the actual correlation. (This is brought on by sensitive dependence to initial conditions.)

Frequency plots of one thousand pseudorandom numbers generated in this fashion will fit a beta distribution whose parameters are: $a = b = \frac{1}{2}$. The same a and b work for any choice of n. The beta distribution is rather obscure, being used mostly for curve fitting and prior distribution in Bayesian statistics. It is lucky that $a = b = \frac{1}{2}$ is one of the few cases in which the cumulative distribution can explicitly be found:

(1) The p.d.f. is:

$$b(x; \tfrac{1}{2}, \tfrac{1}{2}) = \begin{cases} \frac{1}{\pi} \frac{dx}{x-x^2} & , \text{ for } 0 < z < 1 \\ 0 & , \text{ otherwise.} \end{cases}$$

(2) The c.d.f. is:

$$B(x; \tfrac{1}{2}, \tfrac{1}{2}) = \begin{cases} 0 & , \quad \text{if} \quad x \leq 0 \\ \frac{1}{\pi} [\text{Arc sin}(2x - 1) + \frac{\pi}{2}], & \text{if} \quad 0 \leq x \leq 1 \\ 1 & , \quad \text{if} \quad x \geq 1. \end{cases}$$

Using the fact that if X is a random variable with c.d.f. F, then F(X) is a random variable with a uniform distribution on the interval, [0,1], we can transform our pseudorandom numbers into pseudorandom numbers with a uniform distribution on [0,1] by letting $y_i = B(x_i; \frac{1}{2}, \frac{1}{2})$.

Amazingly, the numbers, y_i, are iterates of the tent map! That is the tent map is a theoretical uniform random number generator and the tent map and logistic map are topologically conjugate. (See II.9.) That is, $B \circ f \circ B^{-1} = g$ where B^{-1} is the inverse of the cumulative beta distribution.

The gist of this discussion is that chaotic behavior really is "random". In the eyes of the observer, any sequence of pseudorandom numbers is truly random if the observer doesn't know how to duplicate them, or at least know what algorithm was used to generate them. Thus, randomness is subjective.

BIBLIOGRAPHY

Abraham, Ralph, editor of The Visual Mathematics Library: Dynamics, The Geometry of Behavior, Vols. 1 through 4, Aerial Press, 1985-1989.

Atmanspacher and Scheingraber. "A fundamental link between system theory and statistical mechanics", Foundations of Physics, Vol. 17, No. 9, Sept, 1987.

Bia-Lin, Hao. Chaos, World Scientific Publ., 1985.

Barnsley, Michael and Demko, Stephen. Chaotic Dynamics and Fractals, Academic Press, 1986.

Barnsley, Michael. Fractals Everywhere, Academic Press, 1988.

Beltrami, Edward. Mathematics for Dynamic Modeling, Academic Press, 1987.

Cvitanović, Predrag. Universality in Chaos, Adam Hilger, Ltd., 1984.

Devaney, Robert. An Introduction to Chaotic Dynamical Systems, Benjamin/Cummings Publ., 1986.

Falconer, K. J. The Geometry of Fractal Sets, Cambridge Univ. Press, 1985.

Feigenbaum, Mitchell. "Quantitative universality for a class of nonlinear transformations," Journ. Stat. Physics, Vol. 19, No. 1, 1978.

_____. "The universal metric properties of nonlinear transformations", Journ. Stat. Physics, Vol. 21, No. 6, 1979.

_____. "Universal behavior in nonlinear systems," Los Alamos Science, Summer, 1980.

Frederickson, Paul; Kaplan, James; Yorke, Ellen; and Yorke, James. "The Liapunov dimension of strange attractors," Journ. of Diff. Eqns., Vol. 49, p185-207, 1983.

Glass, Leon and Mackey, Michael. From Clocks to Chaos, the Rhythms of Life, Princeton Univ. Press, 1988.

Gleick, James. Chaos, Making a New Science, Viking Press, 1987.

Hurd, Alan. "Resource letter FR-1: Fractals", Amer. Journ. Phys., Vol. 56, No. 11, Nov, 1988.

Hutchinson, J. E. "Fractals and self-similarity", Indiana University Mathematics Journal, Vol. 30, 1981.

Khinchin, A. I. Mathematical Foundations of Statistical Mechanics, Dover Publ., 1949.

_____. Mathematical Foundations of Information Theory, Dover Publ., 1957.

Mandelbrot, Benoit. The Fractal Geometry of Nature, W. H. Freeman and Co., 1983.

Palmore, Julian and McCauley, Joseph. "Shadowing by computable chaotic orbits", Physics Letters A, Vol. 122, No. 8, June, 1987.

Peitgen, Heinz-Otto and Saupe, Dietmar, editors. The Science of Fractal Images, Springer-Verlag, 1988.

Prigogine, Ilya. From Being to Becoming, W. H. Freeman and Co., 1980.

Rogers, C. Hausdorff Measures, Cambridge Univ. Press, 1970.

Rudin, Walter. Real and Complex Analysis, 2nd ed., McGraw-Hill, 1974.

Willard, Stephen. General Topology, Addison-Wesley Publ., 1970.

Appendix 1:

"Tomorrow's Shapes: The Practical Fractal",
The Economist, 26 Dec 1987, pages 97-101.

Reprinted with permission.

The practical fractal

It is no accident that the inventor of fractal geometry, Dr Benoit Mandelbrot, works for IBM. His new science is a child of the computer age. Without the calculating power to explore its weird avenues, and electronic pictures to fire the imagination, fractal geometry would have remained a mathematical oddity. Instead, it may overtake Euclid

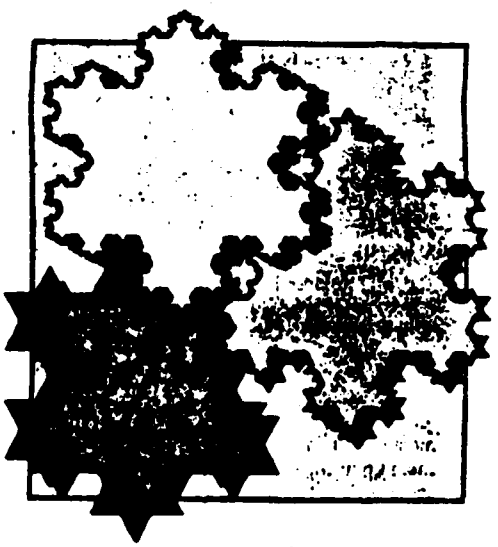
BETWEEN the late 1950s and the early 1970s Dr Mandelbrot invented a branch of mathematics that can describe and analyse the irregularity of the natural world. The key to his theory is a type of shape that he called a fractal. The descriptive power of fractals was soon evident. Fractal forgeries—a type of computer-generated picture—of clouds, mountains and coastlines bear an uncanny resemblance to the real thing. But pretty pictures are not enough to overthrow Euclid. Now, 12 years after Dr Mandelbrot wrote his book, "The Fractal Geometry of Nature", the evidence that fractals can shed light on a wide variety of problems is piling up. The applied fractal has arrived.

Fractals are shapes that look more or less the same on all, or many, scales of magnification. Consider a coastline, the most obvious example of a fractal in nature. Maps of coastlines drawn on different scales all show a similar distribution of bays and headlands. Each bay has its own smaller bays and headlands, *ad* (almost) *infinitum*. The same general structure can be seen in the magnificent sweep of the Gulf of Mexico, the Baie de la Seine, the Pendower Coves near Land's End, the gap between two rocks on the foreshore at Acapulco, and so on down to the individual indentations of a single rock. Coastlines are crinkly however close to them you get.

A mathematical shape that shares this property with coastlines is the Koch snowflake, in which the bays and headlands are successively diminishing equilateral triangles (see diagram). Nature does not sculpt coastlines from triangles, but the Koch snowflake does capture one feature of coastlines well. A tiny piece of coastline, magnified ten times, still looks like a coastline; the same goes for any part of the snowflake. Such objects are said to be "self-similar".

Not so the familiar shapes of old-fashioned geometry, which lose their structure when magnified. For example, the surface of a large sphere appears almost flat when viewed close up, which is why plenty of people used to think the earth itself was flat.

Traditional geometry has to ignore the crinkles, whorls, squiggles and billows of the real world because they are irregular and so do not submit to standard mathematical formulae. The notion of self-similarity lets fractal geometers see a sort of order in the apparent chaos of these shapes. It lets them quantify the roughness and irregularity of a shape and give it a numerical value, known as its fractal dimension.



Dimensions are usually thought of as whole numbers. A line is one-dimensional, a square two-dimensional, a cube three-dimensional. But fractal dimensions are not whole numbers: the Koch snowflake has 1.2618 dimensions, the coastline of Britain has around 1.25 dimensions. The best way to understand this is not to worry about it.

When mathematicians talk about fractal "dimensions", they are not using the term in its ordinary sense. Roughly speaking, if something has more than one but less than two fractal dimensions it is better at filling up space than is an ordinary one-dimensional object (such as a line), but not quite so good as a two-dimensional one (such as a surface). A crinkly line of, say, 1.25 dimensions is better at filling up space than a one-dimensional straight line because you need more ink to draw the crinkle than you do to draw the straight line. A line of 1.26 dimensions is even crinklier and needs even more ink. Some fractal curves are so wiggly and detailed that they fill up nearly all of the surface they are drawn on. So they come within a whisker of qualifying as surfaces—ie, as two-dimensional. That is the crude idea behind fractal dimensions.

Armed with a technique for measuring the irregularity of shapes, the theory of fractals has now been applied to protein structure, acid rain, earthquakes, the fluctuation of exchange rates, oil extraction, epidemics, corrosion, brittle fractures, music, the distribution of galaxies, the level of the Nile and the shapes of clouds, trees, lakes and mountains. Nearly every branch of science studies something that fractals can help with, because all aspects of nature involve some roughness and irregularity.

Superficial science

Begin with surfaces. The shape of surfaces is significant throughout science. When antibodies bind to a virus, or enzymes to a molecule of DNA, they do so because of some affinity for the particular shape of surface involved. Chemical catalysts used in industry work by causing reactions to occur on surfaces. Metallurgists worry about the form of fracture surfaces in metals. Geologists do much the same for mountain ranges. The same shapes may occur on many scales: high-powered microscope pictures of the surface of silicon look rather like the Grand Canyon.

Shapes in nature are usually neither entirely regular nor entirely random. To construct theories of how materials behave, these shapes must be simulated mathematically, either on paper or by computer. Fractals are the perfect tool for the job because they have the right mixture of structure and irregularity. Computer models based on fractals can explore how physical properties of a material vary under different

conditions: how clay flows under pressure, for example.

In 1980 Dr Harvey Stapleton at the University of Illinois at Urbana investigated the magnetic properties of iron-bearing protein molecules. If a crystal is placed in a magnetic field, which is then removed, it loses its magnetic properties in a characteristic fashion. This "relaxation rate" can be quantified; for perfect crystals it is always three. This can be explained mathematically: a crystal is a three-dimensional object and the relaxation rate is equal to the dimension. For proteins, though, Dr Stapleton obtained puzzling relaxation rates, such as 1.7.

He showed that the solution to the puzzle lay in the geometry of the molecules. A typical protein molecule is a long chain of amino acids, folded and crumpled in a most irregular way. The crumpling is fractal—it keeps its structure across a wide range of scales— and the relaxation rate can be computed from its fractal dimension. In fact, the two are equal. So the abstract concept of fractal dimension pays off.

Dr Douglas Rees at the University of California at Los Angeles and his collaborators have shown that protein surfaces—for example, the surfaces of haemoglobin, which transports oxygen in the blood—are fractal. Using computer analysis of the way X-rays are scattered when they hit haemoglobin, they found that the surface of that protein has a fractal dimension of around 2.4. This suggests that the surfaces are very rough, rather like a crumpled paper ball. (If you take a piece of paper the size of this page, crumple it in your fist and then let go, the resulting office-missile has a fractal dimension of 2.5.)

Dr Rees also found that some regions of a protein's surface are smoother—that is, have a smaller fractal dimension—than others. This turns out to be a quite a help for biochemical engineers. Like velcro, proteins stick together best where their surfaces are roughest. And smoother regions seem to be where enzymes, which control the way molecules function, do their work. By measuring the fractal dimension of a protein molecule's surface, the rough can be sorted from the smooth in a precise way. Such techniques could help in the design of synthesised protein molecules for new drugs because they should be able to pinpoint the active sites where enzymes will be able to work.

Sticklable particles

Soot is soft and crumbly because it consists of a loosely knit aggregation of carbon particles. Similar sorts of mucky aggregation are found inside batteries as they corrode, in the process of electroplating, and elsewhere. In 1983, Dr Thomas Witten at Exxon Laboratories in Clinton, New Jersey, and Dr Leonard Sander from the University of Michigan at Ann Arbor found a new way of looking at how such deposits build up, a mathematical model known as Diffusion Limited Aggregation (DLA). In the DLA model, single particles spread out in what mathematicians call a random walk; that is, every so often they move a certain distance in a random direction. They continue this diffusion until they collide with a growing smudge where others have already hit and then stick to it.

Simulations of this process on a computer screen produce branching shapes like irregular ferns, with a fractal dimension of 1.7. Similar random walks in three-dimensional space lead to fractal clusters with a dimension of roughly 2.5. The DLA model has made it possible to analyse and measure many kinds of fractal aggregation.

Dr Jens Feder and his colleagues at the University of Oslo have applied fractals to

Mandelbrot, king of shapes

immunoglobulin clusters. Immunoglobulins are proteins and, like a poached egg, tend to stick together when heated up. But what exactly makes them stick together? Why don't they crumble away? Knowing the fractal dimension of such clusters provides an experimental tool for working out how to build them up in the first place. By simulating the aggregation of particles you can see which processes lead to the right sort of shape. This gives a way to compare theory with experiment. Before fractals were invented, little quantitative work could be done on the way particles pile up as deposits form.

Another process that produces similar branching-tendril shapes is known as viscous fingering. This has been studied for quite some time, but oil companies would like to understand it better. In order to extract oil from a well, water is pumped in under pressure. The oil is then pushed out because oil and water do not mix. But the exact path followed by the water as it flows through the oil is extremely complicated. A better understanding of its swirls and eddies should make it possible to squeeze more oil out of the wells.

The usual way to study this problem of flow uses an apparatus known as a Hele-Shaw cell: two flat glass plates with a thin layer of oil sandwiched in between. Water is fed in through a hole in the middle of one glass plate. At first the water spreads out in a circular disc; but as the disc grows, an inconveniently complicated "dynamic instability" sets in. The boundary between oil and water grows bumps, which in turn grow into "fingers" that penetrate the oil in a star-like pattern. These fingers repeatedly break up in the same way, splitting at the tips when they get too wide. The result is a repeated branching growth rather like a developing plant. It has a fractal dimension of around 1.7.

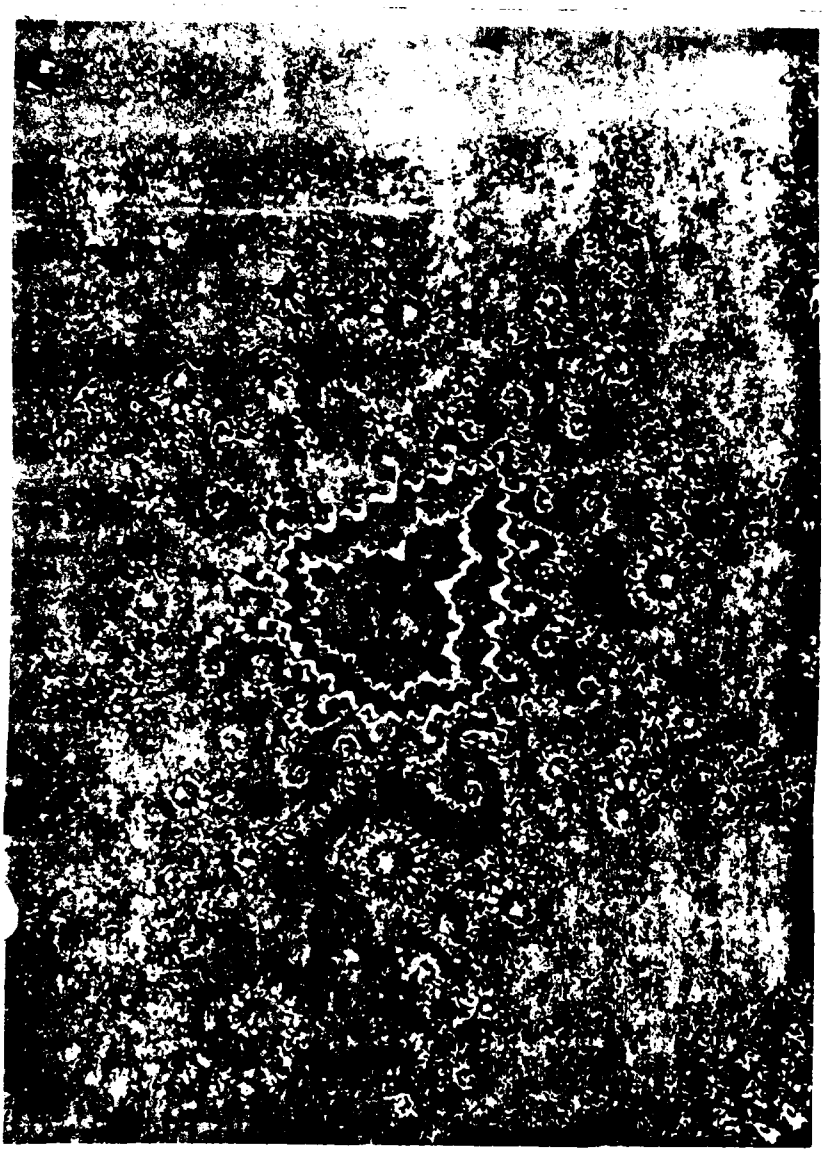
This is remarkably close to the fractal dimension of aggregating particles studied in the DLA model, suggesting that the two processes are related. There is now growing evidence that this guess is right. The mathematics of random diffusion can be recast in a form that is analogous to the mathematics of a fluid moving under pressure. So an experimental coincidence of the fractal dimensions of two processes has led to a surprising theoretical connection between them.

In practice, oil does not occur in large free spaces but is mixed in with particles of rock or sand. Dr Feder and others have investigated viscous fingering in mixtures of rock and sand. They have found that the fractal dimension of the shapes made by the oil as it mixes with water is reduced from 1.7 to about 1.62. The fact that the fractal dimension is lower means that less oil will be squeezed out by the same amount of water when the oil is dispersed in porous rock strata. Knowing this should help companies to extract oil more efficiently by changing their pumping methods to suit the rock strata that occur in a given well.

An epidemic of fractals

Not only particles gather in clusters; people do it, too. Dr Peter Grassberger at the University of Wuppertal in West Germany has used the mathematics of fractal clustering to understand the spread of epidemics. In a not-too-mobile population, a disease that needs close contact between people to pass on spreads rather like the particles in a DLA model. The infection "diffuses" at random through the population and "sticks" where it becomes infectious. A newly infected site acts as a centre for further diffusion.

The process can be simulated on a computer by pretending that individuals live in the cells of a square array, like a huge chessboard, and watching the infection move from one square to another. The effects of different "transmission rules" for the disease (such as that the infection can move only from one cell to an immediate neighbour), various rates of infection (eg, the infection will move to a neighbouring cell only after ten days), and different immunisation proce-



Some mathematicians think that the black snowman shape pictured above—called the Mandelbrot set—may be the most complex object in the universe. It is an infinitely detailed fractal obtained by telling a computer to draw the co-ordinates generated by a mathematical formula. The formula is simple—except for the fact that it involves the square roots of negative numbers. The colours are added later, to taste.

The other pictures on our colour pages are all close-ups of details of the Mandelbrot set. They show that the same whorls, tendrils, spirals and snowmen keep appearing, even when magnified millions of times.

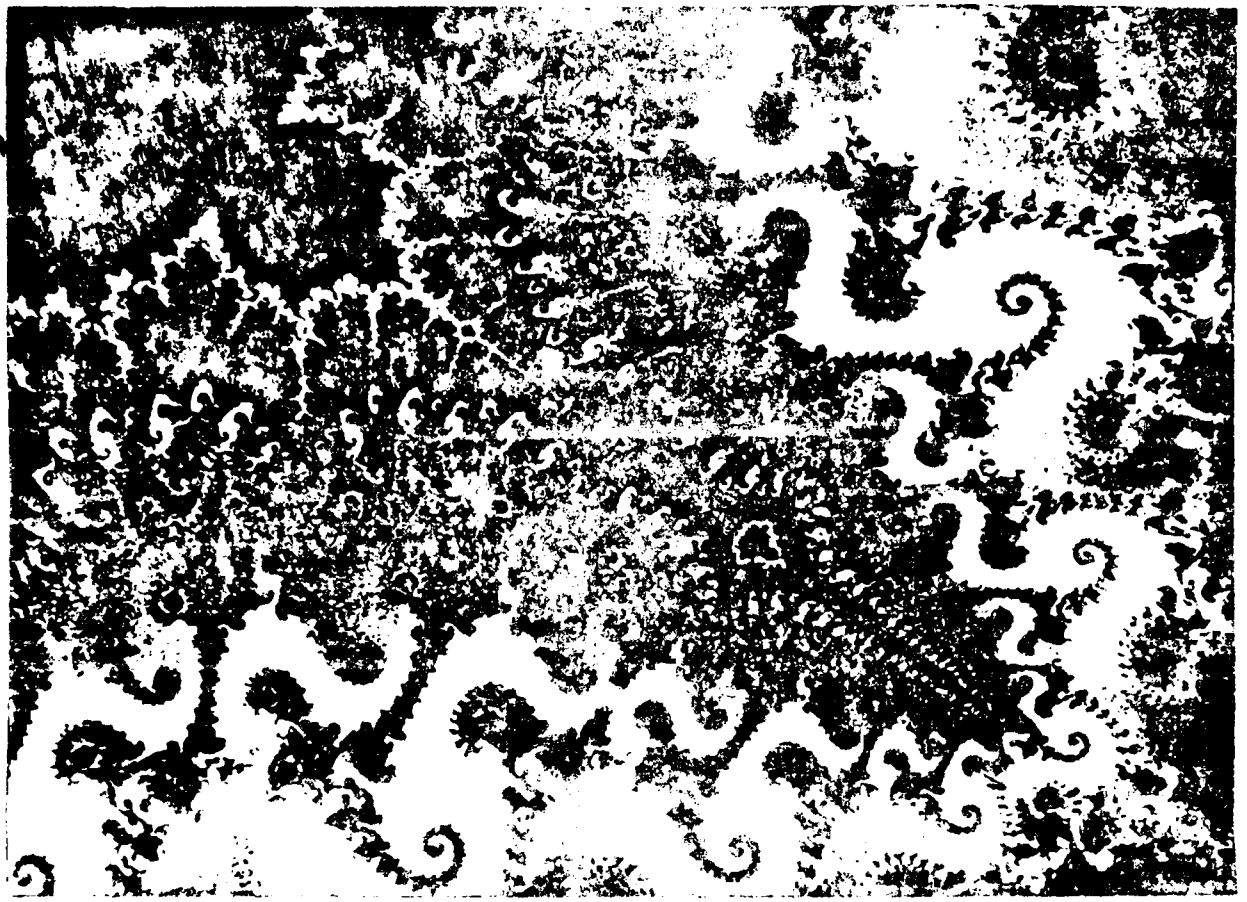
dures can then be explored and tested.

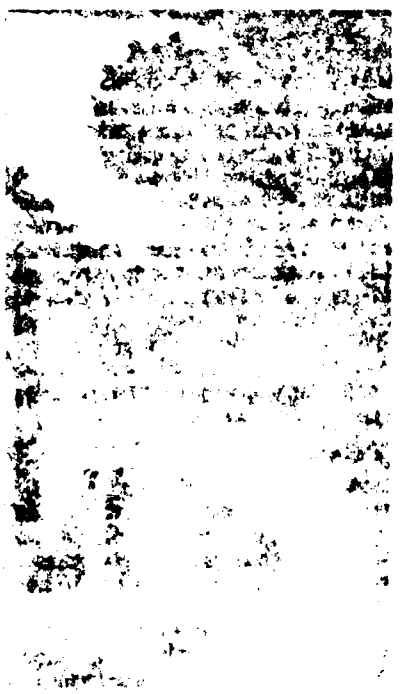
Often the result is a fractal distribution of the disease: the complicated pattern of infected "cells" is the same on several scales. Thus an infection map of a city will look much like the infection map for each block, which will look like the infection map for each street. Clusters of infected cells form, branching in a similar manner to the DLA model. It turns out that the spatial distribution of an infection—where the ill people are in the first place—can be crucial to the way the disease spreads later.

These ideas have implications for the study of AIDS. Simple models take the average rate of infection for a disease and apply it to a uniformly random spread of infected people. But averages do not mean much. Better models recognise that people differ in their behaviour and thus spread the disease at widely varying rates. Average infection rates ignore this variation and so can lead to wildly wrong predictions.

Any model with a fair chance of coming up with the right answers has to recognise that society is an irregular cluster rather than an homogeneous mass. Most of the time people move in their own social circles but different circles can also interact with each other. This makes life horribly complicated for epidemiologists. Dr Robert May, a mathematical biologist at Princeton University, is working on simulations of the spread of AIDS based on ideas in "chaotic dynamics". They deal with the sort of structured irregularities found in fractal geometry.

If the notion of fractal clusters can make the enormous jump from particles to people,





102

...thing goes [illegible]

...Among the [illegible] application [illegible] fractals [illegible] computer graphics. [illegible] produces [illegible] ...pes of its [illegible] computer usually needs enormous amounts of computing power. It would take the combined memory of 10 000 home computers to store the data needed to reconstruct the cratered surface of the moon. This is reasonable enough in astronomical terms, a catalogue of lunar geography, but far too expensive if all you want is a convincing background for a television science-fiction drama. Fractal forgeries that mimic the forms without worrying about precise details can do the job on a relatively modest computer.

One of the most powerful techniques in computer programming is recursion, in which a procedure is broken down into a sequence of repetitions of itself. One normally example is a recursive recipe for building a wall. *First, lay a course of bricks. Then build a wall on top of it.* This is not as silly as it sounds because the instruction "build a wall" can be defined in terms of the first instruction, "*lay a course of bricks*". First you lay a course of bricks, then you lay another on top, and another, and so on. All you have to add is a rule saying when to stop.

Like a recursive set of instructions, fractals break up into copies of themselves. Naturalseeming forms soon start to emerge on the screen when a few random numbers are fed into it to throw up irregular features. By increasing the fractal dimension of the object to be constructed its surface can be made rougher and more pocked. Fractals were used to create the geography of the Moons of Endor in Mr George Lucas's "The Return of the Jedi". Dr Peter Oppenheimer at the New York Institute of Technology uses the fortuitous fractal-like splashing processes to produce lifelike trees and plants.

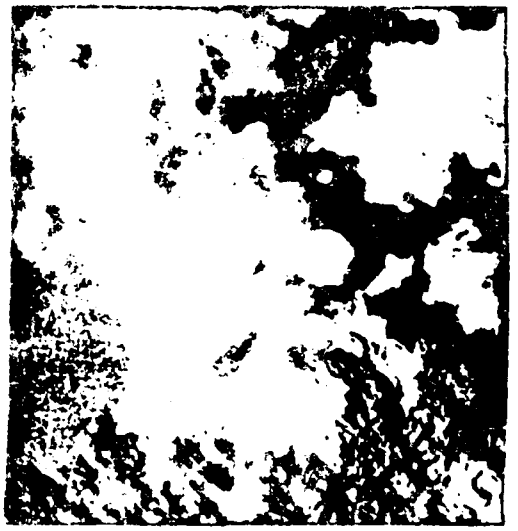
Dr Shaun Lovejoy at McGill University in Montreal has studied real clouds using [illegible] and come to the conclusion that they have the same fractal dimension over seven orders of magnitude—from tens of metres to thousands of kilometres. Such a degree of uniformity is almost unprecedented in the natural world. It means that clouds have no natural length scale. If you were given a photograph of a cloud, with no objects such as trees or houses to tell you the scale, there are no clues in the shape of the cloud itself that will tell you whether it is 1km or 1 000 km across. This was a surprise because the convective loops of air that create clouds led meteorologists to expect a typical scale of around 10km.

Dr Lovejoy has also found that accurate rain maps are [illegible]

[illegible] scales [illegible] size of [illegible] year's rain, for the most part, the same sort of patterns as rain charts for a month, a week, or a single day. So the temporal structure of rain is also fractal.

The idea of distribution across time—ie, the shape drawn on a chart with a time-axis—lets the concept of fractals apply to sound and, more generally to what physicists call "noise". Physicists use the term "noise" to refer to any process that fluctuates irregularly over time, even if it does not involve sound. Different sorts of noise are classified according to their "spectral density". This measures which frequencies occur, and how often. "White" noise has equal amounts of all frequencies (just as white light has equal amounts of all colours) and is entirely random. The noise at any given time cannot be predicted from the noise at earlier times. "Brown noise" is named after a nineteenth century Scottish botanist, Robert Brown, who studied the random motion of tiny particles floating in a liquid. Brown noise is much more ordered than white noise. The noise at any time depends to some extent on what it was in the past. And it contains more high frequencies and fewer low ones than white noise. An intermediate type of noise is known as 1/*f* noise because each frequency, *f*, occurs at a rate that is inversely proportional to its pitch. The higher a note is, the less often it appears. All three types of noise produce wiggly fractal curves if [illegible]. The wiggles have wiggles, and so on.

Remember that anything which changes irregularly over time can be regarded as [illegible] by far the [illegible]



Mythical mountain, conjured-up clouds

[illegible] oceans, the historical variation [illegible] of the Nile, the energy output [illegible] transmission of signals [illegible] can also be found in the behaviour of electronic components.

Music seems to flow like 1/*f* noise. Dr [illegible] at IBM's research laboratories in New York state has analysed variations of pitch in many kinds of music and found that 1/*f* noise predominates. This is equally true of Gregorian chants, Beethoven's symphonies, Debussy's piano works, the rags of Scott Joplin, and the Beatles' Sergeant Pepper album. Only a few modern composers, such as Karlheinz Stockhausen and Elliott Carter, violate this rule. Dr Voss has produced fractal forgeries of music on a computer using white, brown and 1/*f* noise. White music is far too random and brown music far too correlated to sound like any sort of real music. But artificial 1/*f* music, says Dr Voss, sounds as if it is music produced by a foreign culture. He notes that painting, drama and sculpture usually imitate nature in some way. So what does music imitate? He suggests it imitates the 1/*f* noise of the natural world—"the characteristic way our world changes with time".

Fractals are novel in so many ways that it is easy to regard them as wholly isolated from traditional mathematics. That would be a mistake: the theory of fractals is closely linked to at least one apparently independent field, chaotic dynamics. Chaotic dynamics is a belated recognition that purely deterministic—ie, predictable—mathematical models can throw up apparently random results. For example, imagine an insect population that grows from one breeding season to the next according to a fixed numerical formula. The population next year can, in theory, be calculated from this year's. Yet despite such regular laws of growth, the population can fluctuate wildly and unpredictably. This is because tiny errors in the calculation can blow up into wildly divergent predictions over a short time. The result, for all practical purposes, is randomness, or chaos.

Fractals and chaos come together in the study of turbulent flow. Scientists have long been puzzled by the way fluids sometimes flow smoothly and at other times break up into an irregular frothing mass. The same body of fluid can have both turbulent and smooth regions, with a complicated border between them. The classical approach to turbulence sees it as a cascade in which the energy of fluid motion is progressively passed to smaller and smaller vortices. Such a process is fractal because the ever-smaller vortices have the same structure on many scales. The bravest hope for fractals is that they will unravel the mysteries of chaos.

Appendix 2:

"Fractal Applications", Mort La Brecque, MOSAIC,
Winter 86/87, Vol. 17, No. 4, pages 34-38.

Reprinted with permission.

Fractal Applications

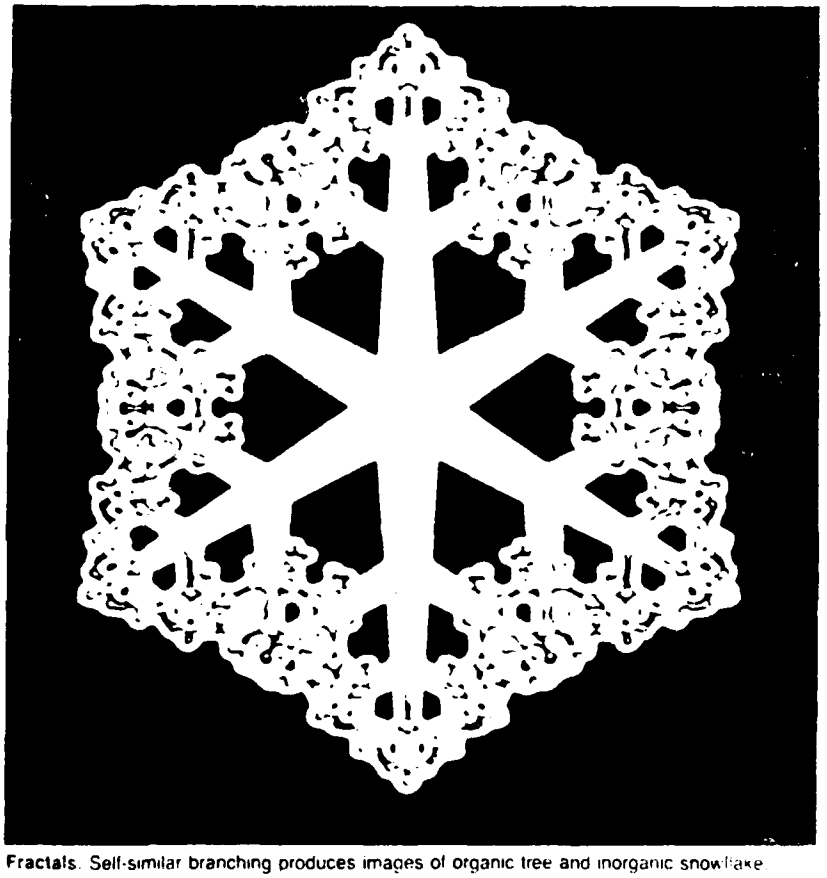
by Mort La Brecque

In 1945, Szolem Mandelbrojt, a prominent French mathematician whose specialty was the abstruse field of complex analysis, had some words of advice for a gifted nephew. It's very nice for children to get some feel for geometry, Mandelbrojt counseled the young man, who was about to undertake his higher mathematical education, but you must leave it behind; a mature mathematician does not use visual images.

Like many other youths before and since, Benoit Mandelbrot decided to reject the wisdom of his elders. He had developed a passionate attachment to shape and form that he could not relinquish. The decision, in this case, has been vindicated by time. Not only has his pursuit of a geometrical grail guided Mandelbrot to extraordinary success as a mathematician, but the fractal geometry to which it led has profoundly influenced contemporary science as well.

The concept with which Mandelbrot's name has become synonymous is deceptively simple: An object that is self-similar has a rough shape to one degree or another made of parts which, when magnified, resemble the whole. The parts are divisible into look-alike entities—and this geometric cloning continues forever, at least in the abstract world of pure mathematics. Mandelbrot coined the word fractal to describe a self-similar object—the Latin *fractus*, meaning irregular and fragmented, suits its twisted and tormented form. (See "Fractal Symmetry" by Mort La Brecque, *Mosaic* Volume 16 Number 1.)

If fractal is a new word, self-similarity can hardly be considered a new idea. The image of a flea upon a flea that Swift employed in *On Poetry* refers to the London literary scene but was inspired by Gottfried Wilhelm von Leibniz, the seventeenth- and eighteenth-century German philosopher and mathematician. Although Leibniz's exuberant optimism—all is for the best in this best of all possible worlds—was wickedly satirized by Voltaire in *Candide*, he also had



Fractals. Self-similar branching produces images of organic tree and inorganic snowflake.

> So, Nat'ralists observe, a Flea
> Hath smaller Fleas that on him prey;
> And these have smaller Fleas to bite 'em;
> And so proceed *ad infinitum*.
>
> —Jonathan Swift,
> *On Poetry. A Rhapsody*

This is part one of a two-part special report by Mort La Brecque on fractal applications.

Fractals are far more than the fantastic fruits of the crossmatching of geometric theory and computer graphics. Both the spawn and the seed of a mathematical revolution, they are influencing an increasing range of scientific undertakings.

Peitgen. There is so much complexity in these systems; so much wildness and structure.

the worthier notion that in every drop of dew is an entire world with its own drops of dew, themselves containing worlds—and dewdrops—of their own.

"That has a strong, almost theological resonance for man," says Mandelbrot. Fractals have become popular, he contends, because of the intrinsic appeal they have for us, now that their nature is understood.

Science has succumbed to the lure of fractals for less mystical reasons. "The mathematical concepts related to fractals have gone a long way in unifying areas of physics, chemistry, and biology that were previously obscure and couldn't be approached," says chemist Raoul Kopelman, who works at the University of Michigan.

The fractal world

Mandelbrot, who has used the concept of self-similarity since the late 1950s, believed initially that he had hit upon a basic organizing principle of nature. "There was a very widespread feeling, fostered by many poets and great writers, that nature has an organic complication which no mathematics can ever imitate," he says. "It's ironical that fractals, many of which were invented [by nineteenth-century mathematicians] as examples of pathological behavior, turn out not to be pathological at all. In fact, they are almost the rule in the universe. Shapes which are not fractal are the exception."

Those exceptional shapes are the perfect lines, planes, and cubes of Euclid, which have been part of our culture for 2,000 years. "I love euclidean geometry," he says, "but it is quite clear that it does not give a reasonable presentation of the world. Mountains are not cones, clouds are not spheres, trees are not cylinders. Almost everything around us is essentially noneuclidean."

There is both an upper and a lower limit, however, to the size range over which natural fractals are fractal: At certain points, they may either become smooth or rough but not fractal, or else they reach their particular ultimate state in bigness or smallness. Moreover, natural fractals are fractal in a statistical or stochastic sense, a particular shape giving no clue to the length scale at which it was determined and not looking exactly like a shape on a different length scale. Exactitudes and infinities exist only in the province of mathematical fractals.

The range of natural phenomena encompassed by self-similarity is astonishing. In addition to the mountains, clouds, and trees mentioned by Mandelbrot and the galactic clusters and turbulent flows on which so much of his work centered, there are proteins, acid rain, deposition, the surface of the earth, fault zones, earthquake patterns, [illegible] mechanical and electrical systems, chemical reactions—even the pattern in which oil and water do not mix. Fractal materials also include amorphous materials like glass, colloidal aggregates, electrodeposited metals, electrolytes, thin films, coal, and ceramics. Even the cracks in ceramics are actually fractals.

An equally impressive list can be culled from the world of mathematics, from whence the shape of fractals sprang. To the Cantor bar, Peano curve, and Sierpinski gasket of the late nineteenth and early twentieth centuries—now often used by physical and life scientists as models of natural fractals—has been added the Mandelbrot set of the late twentieth century. This bizarre object, simultaneously well-ordered and wildly chaotic, is the focus of intensive scrutiny by some of the best mathematical minds.

Mandelbrot

All the activity in science and mathematics that has identified and explicated those fractals can be traced from a burgeoning number of younger acolytes to the central avuncular figure of Benoit Mandelbrot. He is to fractal geometry what Einstein was to relativity and Freud to psychoanalysis. Although he has often ruffled the feathers of some colleagues by his immodest insistence that credit go where credit is due, most would probably agree that his peripatetic imagination, proselytizing fervor, and sheer dogged persistence virtually created the field.

Mandelbrot, now at Harvard University, conducted most of his work over a period of nearly 30 years at IBM's Thomas J. Watson Research Center, where he continues as an IBM fellow. Oddly enough, his interest in fractals began about that time, when he was studying short- and long-term commodity price changes. The structure of the fluctuations, he discovered, could be reproduced by a self-similar forgery. (Were he to focus on one field today, he says, it would be economics.)

There followed work on a sequence of problems that was distinguished, outside of the fractal connection, by a total absence of relatedness: errors in the transmission of data over telephone channels, the widespread phenomenon called 1/f noise, and fluctuations in the level of the Nile River. Concurrently, Mandelbrot was developing his knowledge of mathematics and creating new

---

La Breque has contributed to Mosaic on many subjects. His most recent article was "Many-body Problem," which appeared in [illegible]

structures that he would then apply to the scientific problems that had confounded him. The oscillation between different fields of science and mathematics has been constant, a hallmark of his entire career.

Beginning in 1964, Mandelbrot began to consolidate his findings from his earlier disparate studies, at the same time adding to their number. He also recalls the ensuing decade as one in which he met great resistance from the scientific establishment. "I was certainly the only person doing these things," he says, "except for friends who occasionally joined me because they were interested in a particular project."

The turning point apparently came in 1975, with the publication of his first book in French, translated into English in 1977 as *Fractals: Form, Chance, and Dimension*. In the late 1970s and early 1980s, his work was finally adopted by the physics community, first by physicists newly engaged in studies of the chaotic behavior born of turbulence and then by statistical physicists, a larger group interested in a broader range of phenomena. Those influential converts and the publication of Mandelbrot's best-selling second book, *The Fractal Geometry of Nature*, in 1982, brought chemists, biologists, computer scientists, geophysicists, astrophysicists, materials scientists, meteorologists, mechanical engineers, and scientists from other disciplines into the fractal fold.

"The number of people involved is becoming enormous," says Mandelbrot. "I don't know all of them, and I can't read everything they write." Although he organized the first scientific meetings on fractals by himself—he was the only person knowledgeable enough to do so—such arrangements must now be left largely to others.

Devaney. Application of Julia sets must wait.

Computer graphics

Mandelbrot likes to point out that his own labor was enhanced immeasurably by a stroke of good luck—the simultaneous development of computer graphics that was to prove invaluable in the application of fractals to science. "My first work had no pictures whatsoever, and I found I couldn't make my ideas understood by my audience," he says. "They thought I was making a fine technical distinction that didn't truly matter to the central issues of their fields," he explains.

In the late 1960s, he realized that he could use a simple pen writer to draw real records of river-level fluctuations side by side with fractal forgeries. The illustrations convinced a hydrologist that Mandelbrot was making a point of fundamental significance. Shortly afterwards, Mandelbrot acquired access to some of the first computers for making graphics and, using them to construct fractal forgeries of mountains, fooled people who saw photographs and a motion picture of the images into thinking that they were the real thing. Since then, he has upgraded both his graphics and photographic equipment to produce, with IBM colleague Richard Voss, even more natural, counterfeit mountainous terrains.

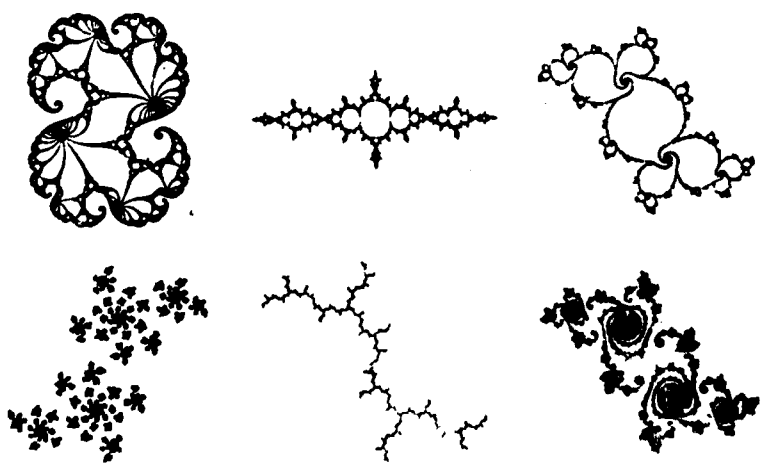
Computer graphics proved to be more than a rhetoric Mandelbrot could use to explain his thoughts to others. The pictures acquired a life of their own, stimulating him to make new conjectures and to advance his own research.

"For me, the most important instrument of thought is the eye," he says. "It sees similarities even before a formula has been created to identify them." The use of computer graphics as an intuition-builder has persuaded Benoit Mandelbrot and others that technology can be as great an influence on science as science is on technology.

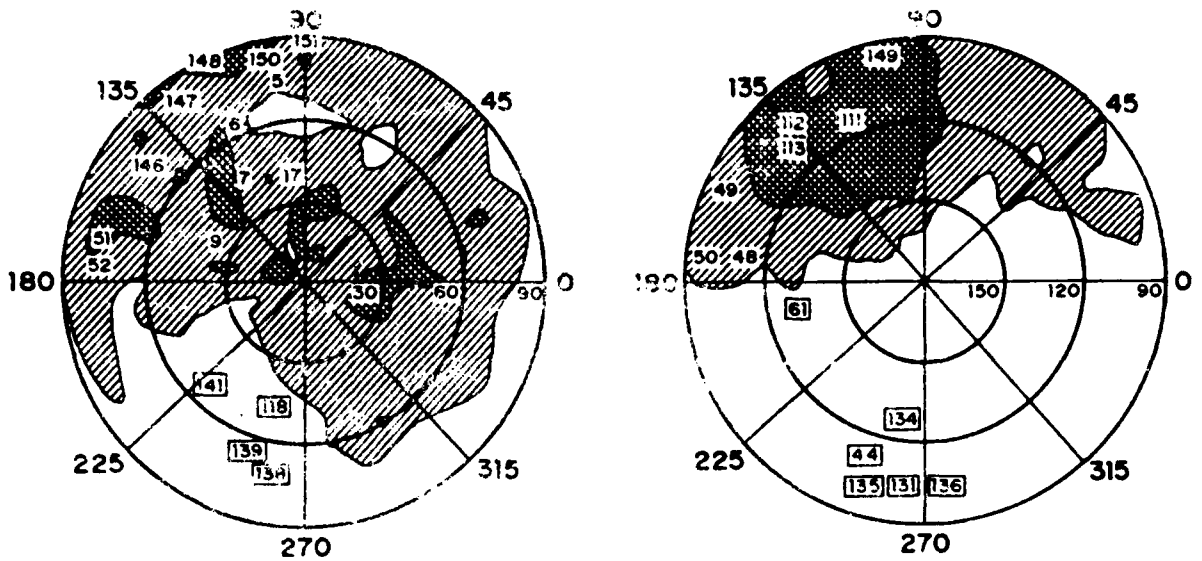
Image synthesis

The application of computer graphics is vast, and in the realm of image synthesis also beginning to have an influence on art, both representational and abstract, displayed in forms both traditional and technological. Some of the most striking computer-generated pictures there are have been made with fractal algorithms.

The first images made with a computer, created by Ivan Sutherland at MIT in the late 1950s, were composed first of dots and then of lines. By the late 1960s, researchers at the University of Utah had depicted objects made of triangles and polygons. The French mathematician Pierre Bezier, one of the founders of computer-aided design, sculpted curved surfaces in the early 1970s, but those reference shapes were merely drawings, highly inaccurate for the needs of the airplane and automobile industries, where they were being used.



Julia sets: "A [illegible] with which [illegible] teeter on the brink of [illegible] their states



Enzyme fractals. Display based on x-ray crystallography shows surface structures of two molecules of the enzyme superoxide dismutase.

An improved method for making curved surfaces by splitting them into polygons was thereafter perfected by Loren Carpenter, then of Boeing Computer Services.

Although this explicit method has become the standard for making curved surfaces, it leaves a lot to be desired. The entire image must be done by hand, that is with direct mental intervention of the human model-builder, who must specify—explicitly—the coordinates of every corner or control point that lies in a curve. Since any complex shape, such as a tree or a hill, could require hundreds of thousands or even a million points to be described adequately, the cost in human labor is prohibitive, except for the crudest approximations.

In 1978, Carpenter, reading a review of Mandelbrot's first book, saw immediately that fractals offered a way of overcoming those limitations and began to develop an algorithm to do so. That idea occurred at the same time, independently, to Alain Fournier and Don Fussell of the University of Texas at Austin, and eventually the three published a paper together in a journal of the Association for Computing Machinery.

Computers take over from the human—for the most part—in their stochastic modeling techniques. "You let the computer generate random shapes that are constrained by certain rules," Carpenter explains, "and if you use rules that are based on abstracted from processes that yield natural forms, you can assure their appearance."

To produce a simple fractal curve that might be used to form a fractal mountain, for example, the computer program vertically breaks a line somewhere between the two end points. That distance is constrained by three rules determined by the programmer. It must be proportional to the length of the distance between the end points so that the same approximate shape is produced at all length scales; it must have a global scale factor that specifies the degree of roughness desired in the curve; it must have a certain distribution of randomness, usually that of a Gaussian curve on which most of the points are clustered near the center.

Loren Carpenter. Art decidedly not for art's sake.

Once the line is broken, the two resulting lines are severed again according to the same principles, as are the four lines that result from the second operation. This recursive splitting is repeated until a crinkly curve is produced of such statistical complexity that a human being could never reproduce it without a computer. More important, the mountain assembled from such curves will look like a real mountain.

Filmed fractals

Carpenter created a two-minute-long animated film to illustrate his ideas. Shown at the 1980 ACM SIGGRAPH—Special Interest Group in Graphics—conference "Vol Libre" portrayed a simulated flight over a fractal landscape, which included a number of different fractal processes. It also demonstrated that computer-animated film could be entertaining as well as instructive.

The movie apparently succeeded beyond Carpenter's wildest dreams. Talent scouts from George Lucas's Lucasfilms, who attended the conference, were sufficiently impressed to hire Carpenter for Pixar, the company's fledgling computer graphics division. (Its original mandate to computerize im-

P 35

This is the last Page
included in Appendix 2
See Appendix "2"

Appendix 3:

"Classical Chaos", Roderick V. Jensen, American Scientist, March/April 1987, Vol. 75, pages 168-181.

Reprinted with permission.

Classical Chaos

Roderick V. Jensen

A wide variety of natural phenomena exhibit complicated, unpredictable, and seemingly random behavior. Common examples include the turbulent flow of a mountain stream, the changing weather, and the swirling patterns of cream, slowly stirred, in a cup of coffee. The paradigm for this class of macroscopic phenomena is the problem of turbulent flow in fluids (Fig. 1). Additional examples of complex, irregular behavior occur in the dynamics of molecules and atoms in a gas or charged particles in a plasma. These microscopic systems define another class of important physical problems which raise a disturbing question: How can the deterministic and reversible motions of individual particles give rise to the irreversible behavior of the system, as described by statistical mechanics and thermodynamics?

New methods for studying chaotic behavior make the unpredictable more understandable but also raise disturbing fundamental questions

Although physics has made monumental strides in the last hundred years, theoretical descriptions of these complex phenomena have remained outstanding unsolved problems. The difficulty lies in the nonlinear character of the mathematical equations which model the physical systems: the Navier-Stokes equations for fluid flows and Newton's equations for three or more interacting particles. Since these equations do not generally admit closed-form analytical solutions, it has proved extremely difficult to construct useful theories that would predict, for example, the drag on the wing of an airplane or the range of validity of statistical mechanics. However, in the last ten years considerable progress has been made, using a unique synthesis of numerical simulation and analytical approximation.

The key to the recent progress has been the use of high-speed digital computers. In particular, high-resolution computer graphics have enabled the "experimental" mathematician to identify and explore ordered patterns which would otherwise be buried in reams of computer output. In many cases the persistence of order in irregular behavior was totally unexpected; the discovery of these regularities has led to the development of new analytical methods and approximations which have improved our understanding of complex nonlinear phenomena.

This novel approach, which combines numerical "experiments" with mathematical analysis, has given rise to a new interdisciplinary field called nonlinear dynamics. The work done in this field has been applied not only to problems in physics but also to a wide variety of nonlinear problems in other scientific fields, such as the evolution of chemical reactions (*1*), the feedback control of electrical circuits (*1*), the interaction of biological populations (*2*), the response of cardiac cells to electrical impulses (*3*), the rise and fall of economic prices (*4*), and the buildup of armaments in competing nations (*5*). In this article I will limit myself primarily to physical problems. However, I hope that readers will recognize the applicability of these methods to their varied fields, since the difficulties in solving nonlinear equations are common to every branch of science.

Nonlinear dynamicists use the word "chaos" as a technical term with a precise mathematical meaning to refer to the irregular, unpredictable behavior of deterministic, nonlinear systems (*6*). Contrary to what Isaac Newton may have believed, the deterministic equations of classical mechanics do not imply a regular, ordered universe. Although most modern physicists and gamblers would concede that dynamical systems with large numbers of degrees of freedom, such as the atmosphere or a roulette wheel, can exhibit random behavior for all practical purposes, the real surprise is that deterministic systems with only one or two degrees of freedom can be just as chaotic.

Traditionally, the fundamental problems associated with the origins of chaos in turbulent flows, the microscopic foundations of statistical mechanics, and the appearance of random behavior in a variety of other fields have been avoided by using the argument that so many particles and degrees of freedom are involved that it would not be humanly possible to describe these

Roderick V. Jensen is an associate professor of applied physics at Yale University. He is a graduate of Princeton University (A.B. in physics 1976, Ph.D. in astrophysical sciences 1981), where his dissertation research was devoted to the statistical description of chaotic dynamical systems with applications to plasma physics. His current research is concerned with the role of chaos in the foundations of statistical mechanics and the investigation of chaotic behavior in quantum systems. This work is supported by an Alfred P. Sloan Fellowship and a Presidential Young Investigator Award from the National Science Foundation. Address: Mason Laboratory, Department of Applied Physics, Yale University, Yale Station, New Haven, CT 06520.

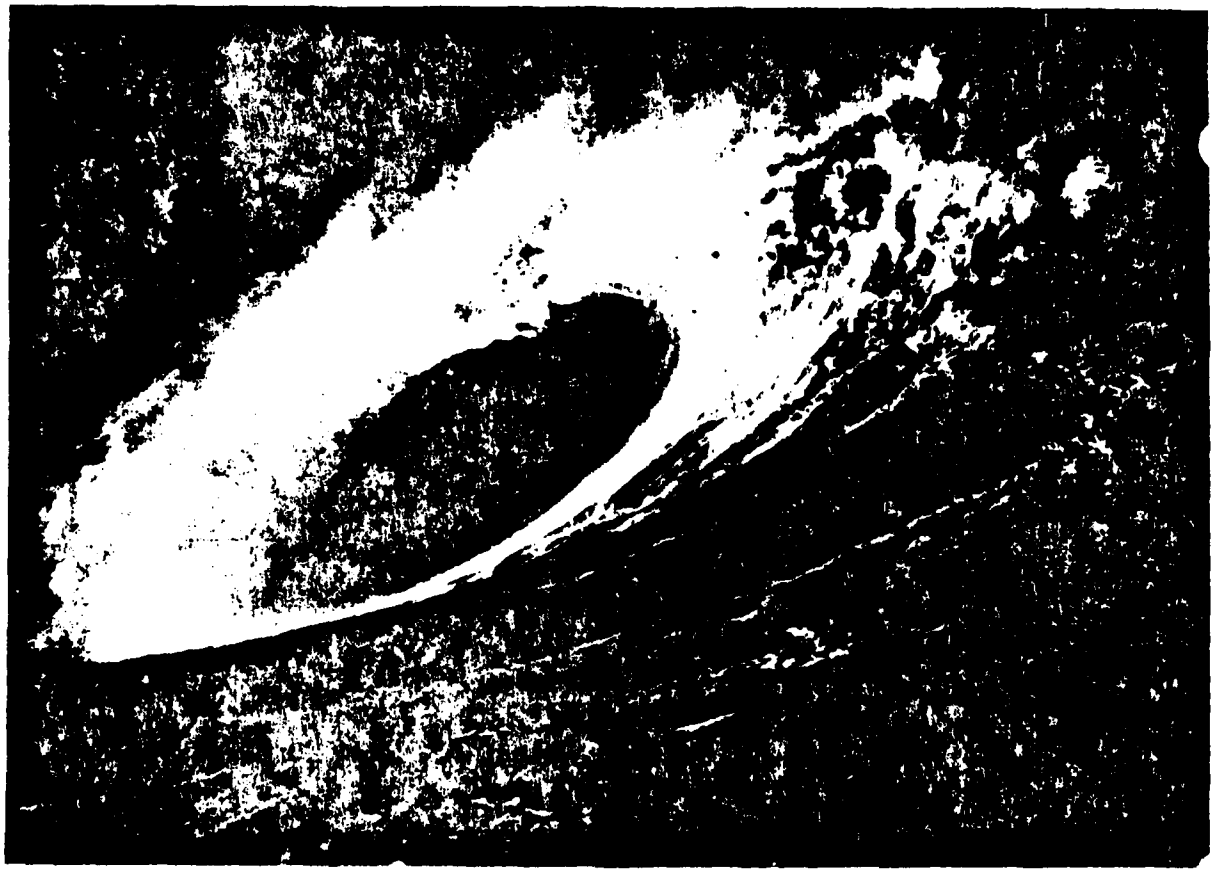


Figure 1. When motion becomes chaotic, the results are unpredictable and sometimes disastrous. In classical dynamics, the behavior of turbulent fluids has proved extremely difficult to predict—as we know, for example, from weather forecasting. But new insights about the nature of chaos have revealed an underlying structure that is common in many natural systems and even in human social behavior. These insights have been applied to such problems as the evolution of chemical reactions, the control of electrical circuits, the growth of biological populations, the response of cardiac cells to electrical impulses, the rise and fall of economic prices, and the buildup of armaments. (Photograph © Four By Five.)

phenomena from first principles. However, the discovery of much simpler systems which can nevertheless exhibit behavior as complicated as these standard examples means that we no longer have to throw up our hands in despair. Using the computer as a laboratory apparatus to study these simple systems, we can begin to explore and understand chaotic, irregular, and unpredictable phenomena in nature.

In this review I will concentrate on phenomena which are well described by classical physics and, consequently, on problems of "classical chaos." Unfortunately, the question of chaos in quantum physics remains controversial. At present, "quantum chaos" is a poorly characterized disease for which we have only identified some of the possible symptoms. Both an unambiguous definition as well as the very existence of quantum chaos remain open problems. In contrast, we have a clear understanding of the symptoms and causes of classical chaos, if only a partial understanding of the cure.

I will start by examining in detail two deceptively simple nonlinear dynamical systems which exhibit a transition from regular, ordered behavior to chaos. These examples will graphically illustrate the irregular, unpredictable, but nevertheless deterministic behavior we call chaos. Then, after formulating a precise definition of classical chaos, I will attempt to dispel the longstanding psychological prejudice which insists on a distinction between deterministic and random behavior by showing that the chaotic behavior of deterministic dynamical systems can be indistinguishable from a random process.

This deeper understanding of chaos will lead, finally, to a slightly more philosophical discussion of where classical chaos really comes from and what it is good for. We will see that investigations of nonlinear dynamical systems have suggested partial answers to some of the fundamental problems of turbulence and statistical mechanics which were first formulated in the nineteenth century. However, this research has also raised new questions, more profound than those they have answered, relating to twentieth-century problems arising from Gödel's incompleteness theorem, the theories of algorithmic and computational complexity of modern computer science, and the principles of quantum mechanics (7).

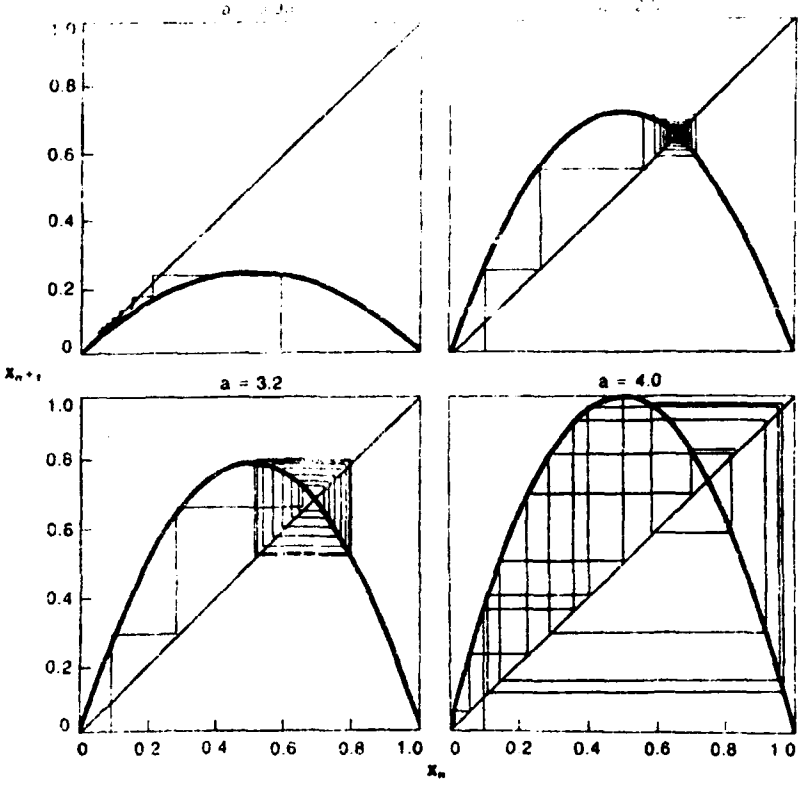


Figure 2. Many dynamical systems can be approximated by the logistic map (equation 1), which then predicts, for example, the size of a changing biological population, the fluctuation of economic prices, or the dynamics of a periodically kicked and damped nonlinear oscillator. Equation 1 defines an inverted parabola, plotted here for four different values of a. Once an initial value, x_0, is specified, the evolution of the system is fully determined. One can find the values of x at succeeding time-steps by tracing the colored lines on the appropriate graph: from x_0 vertically to the parabola for x_1, then horizontally to the 45° line and vertically back to the parabola for x_2, and so on for succeeding values of x_{n+1}. When a is less than 1, as in the graph at the upper left, all initial conditions converge to 0 (the population becomes extinct, the price falls to zero, etc.). When a is increased to a value between 1 and 3, as in the upper right, almost all initial conditions are attracted to a fixed point. When a is larger than 3, however, the fixed point becomes unstable; at $a = 3.2$ there are two fixed points between which the value for x eventually oscillates. As a continues to increase, there can be more and more fixed points, and for many values of a, as when $a = 4$, the values for x wander over entire intervals in an apparently random fashion.

Examples of chaotic dynamical systems: The logistic map

Perhaps the simplest example of a nonlinear dynamical system is the celebrated logistic map. This system is described by a single difference equation

$$x_{n+1} = ax_n(1 - x_n) \qquad (1)$$

which determines the future value of the variable x_{n+1} at time-step $n + 1$ from the past value at time-step n. The time-evolution of x_n generated by this single algebraic equation exhibits an extraordinary transformation from order to chaos as the parameter a, which measures the strength of the nonlinearity, is increased.

Although nonlinear difference equations of this type have been studied extensively as simple models for turbulence in fluids, they also arise naturally in the study of the evolution of biological populations. In fact, the review article on the logistic map by the biologist Robert May (2) is a historical milestone in the modern development of nonlinear dynamics. Therefore, for illustrative purposes we will examine the use of the logistic map as a crude model for the annual evolution of a single biological population, x_n, say that of gypsy moths in the northeastern United States, which exhibits wild and unpredictable fluctuations from year to year. However, we could equally well consider the evolution of economic prices determined by a nonlinear "cobweb" model with nonmonotonic, backward-bending supply and demand curves (4) or the dynamics of a periodically kicked and damped nonlinear oscillator.

Writing equation 1 in a slightly different form, $x_{n+1} = ax_n - ax_n^2$, we see that it is a simple quadratic equation, with the first term linear and the second term nonlinear. When the original population x_0 is small (much less than 1 on a normalized scale, where 1 might stand for any number, such as 1 million individuals), the nonlinear term can initially be neglected. Then the population at time-step (year) $n = 1$ will be approximately equal to ax_0. If $a > 1$, the population increases. If $a < 1$, the population decreases. Therefore, the linear term in equation 1 can be interpreted as a linear growth or death rate which by itself would lead to exponential population growth or decay. If $a > 1$, the population will eventually grow to a value large enough for the nonlinear term, $-ax_n^2$, to become important. Since this term is negative, it represents a nonlinear death rate which dominates when the population becomes too large. Biologically, this nonlinear death rate could be due to the depletion of food supplies or the outbreak of diseases in an overcrowded environment.

As emphasized in May's review article, the dynamics of this map and the dependence on the parameter a, which measures the rate of linear growth and the size of the nonlinear term, are best understood using a graphic analysis. Consider the graphs of x_{n+1} versus x_n (called "return maps") displayed in Figure 2 for four different values of a. Equation 1 defines an inverted parabola with intercepts at $x_n = 0$ and 1 and a maximum value of $x_{n+1} = a/4$ at $x_n = 0.5$. Using these return maps, we can get a qualitative understanding of the dynamics of the logistic map without performing any calculations. The successive values of the populations can be determined simply by tracing lines on these graphs. Just start your pencil at an initial x_0 and move vertically to the parabola to get x_1. At this point you could return to the horizontal

axis to repeat this procedure using the value of x_1 to get x_2, but it is more convenient simply to trace horizontally to the 45° line and then vertically to the parabola again, as shown by the colored lines in each graph.

This graphic analysis tells us that if the normalized population starts out larger than 1, then it immediately goes negative, becoming extinct in one time-step. Moreover, if $a > 4$, the peak of the parabola will exceed 1, which makes it possible for initial populations near 0.5 to become extinct in two time-steps. Therefore, we will restrict our analysis to values of a between 0 and 4 and to values of x_0 between 0 and 1.

For values of $a < 1$, the population always decreases to 0, as shown for $a = 0.95$ in Figure 2. The intersection of the parabola with the 45° line at $x_n = 0$ represents a stable fixed point on the map. Because a is small, perturbation theory can be used to verify that almost all initial populations are attracted to this fixed point and become extinct. However, for $a > 1$ this fixed point becomes unstable. (This is readily verified by tracing the dynamics in the second graph or by applying a local perturbation theory for small populations.) Instead, the parabola now intersects the 45° line at $x = (a - 1)/a$, which corresponds to a new fixed point. Conventional perturbation theory gives no hint of the existence of this nonvanishing steady state population.

For values of a between 1 and 3 almost all initial populations evolve to this equilibrium population. Then, as a is increased between 3 and 4, the dynamics change in remarkable ways. First, the fixed point becomes unstable and the population evolves to a dynamic steady state in which it alternates between a large and a small population. A time-sequence converging to such a period-2 cycle is displayed in Figure 2 for $a = 3.2$: the population eventually cycles between two points on the parabola, $x_n \sim 0.5$ and $x_n \sim 0.8$, in alternate years. For somewhat larger values of a this period-2 cycle becomes unstable and is replaced by a period-4 cycle in which the population alternates high-low, returning to its original value every four time-steps. As a is increased, the long-time motion converges to period-8, -16, -32, -64,... cycles, finally accumulating to a cycle of infinte period for $a = a_{inf} \sim 3.57$.

This sequence of "period-doubling bifurcations" in the long-time, steady state behavior of the logistic map is clearly displayed in Figure 3. The graph shows the steady state values of the population as a function of a between 3.5 and 4. For $a < 3$ only a single steady state value of $x = (a - 1)/a$ would be displayed. For $a > 3$, we get two steady state values, then four, then eight, and so on. Each bifurcation in Figure 3 thus represents a doubling of the number of steady state values and a doubling of the time-steps in a period.

The range of a over which a single cycle is stable decreases rapidly as the period of the cycle increases, which accounts for the rapid accumulation of cycles with larger and larger periods. In fact, having observed this period-doubling sequence in numerical experiments, Feigenbaum was able to prove, using a remarkable application of the renormalization group, that the intervals over which a cycle is stable decrease at a geometric rate of ~ 4.6692016. The tremendous significance of this work is that this rate and other properties of the period-doubling sequence are universal in the sense that they appear in the dynamics of any system which can be approximately modeled by a nonlinear map with a quadratic extremum (*8*). Feigenbaum's theory has subsequently been confirmed in a wide variety of physical systems such as turbulent fluids, oscillating chemical reactions, nonlinear electrical circuits, and ring lasers (*1*).

The investigation of period doubling in nonlinear dynamical systems provides a superb example of the interplay between numerical "experiments" and analytical theory. However, this sequence of regular periodic orbits is only the precursor to chaos. Since the period-doubling route to chaos has been the subject of several other review articles and texts (*2, 8–10*), I will now move on to still larger values of a, where the dynamics of the logistic map are truly chaotic.

For many, if not most, values of $a > 3.57...$ the bifurcation diagram shows that the long-time behavior of the population is aperiodic and ranges over continuous intervals of x. As I will demonstrate, the evolution of populations in these continuous intervals is indistinguishable from a random process, even though the

Contrary to what Isaac Newton may have believed, the deterministic equations of classical mechanics do not imply a regular, ordered universe

logistic map is fully deterministic in the sense that there are no "random" forces and the future is completely determined by the initial condition, x_0.

However, we also find windows of periodic behavior embedded in this chaotic regime. The most prominent window corresponds to a period-3 cycle for $a \sim 3.83$, in which the population increases in two successive years and decreases in the third. Moreover, as a is increased within this window of stability, the period-3 cycle can also be seen to exhibit period-doubling bifurcations to period-6, -12, -24, ... cycles. In fact, between a_{inf} and $a \sim 3.83$ there are windows of stability for every integer period, which terminate in a period-doubling cascade back to chaos. Although the windows of stability for most of the higher-order cycles are too narrow to be seen in Figure 3, a period-5 and a period-6 cycle can be readily discerned.

It is a remarkable mathematical fact that, although these intervals of stability are dense throughout the range of a, it is not correct to conclude that the set of values of a for which the motion is truly chaotic is negligibly small. On the contrary, this set has been proved to have a nonvanishing measure (*11*). In other words, if the exact evolution of x_n looks chaotic, then it probably is; we are not necessarily being deceived by a very long, but periodic, cycle. In particular, the irregular dynamics for $a = 4$, which deterministically spans the entire unit interval, is easily shown to meet the definitions of both a chaotic and a random process formulated later in this article.

Another striking feature in the bifurcation diagram is the dark streaks which mark the upper and lower boundaries, and criss-cross the chaotic domain. The dark

streaks represent values of x which are more probable and visited more often during the chaotic evolution. These ordered structures were discovered "experimentally" in high-resolution graphs, like Figure 3, displaying hundreds of thousands if not millions of iterations of the logistic map. Once discovered, their explanation was found to be simple (*12*). The streaks are located at the future values of the "critical" population, $x_0 = 0.5$. The upper bound of values for x is determined by the heights of the inverted parabolas, $x_1 = a/4$, as diagrammed in Figure 2, and the lower bound and all the interior streaks in Figure 3 by the subsequent iterates. The reason that populations have a higher probability of passing through values near the trajectory of $x_0 = 0.5$ is that the slopes of the parabolas on the return maps (Fig. 2) vanish there, which tends to compress nearby trajectories. Moreover, the intersections of these dark streaks in Figure 3 correspond to "crises" in the chaotic dynamics, where disjoint intervals of chaotic orbits collide to form larger regions, and they have been a topic of recent research (*13*). The most spectacular crisis is readily visible at $a \sim 3.68$.

High-resolution computer graphics have enabled mathematicians to identify ordered patterns which would otherwise be buried in reams of computer output

The discovery and explanation of such regular structures in the chaotic domain is not just an amusing exercise for experimental mathematicians; rather, an understanding of these probability distributions has important practical applications. Since an analytical description of the chaotic evolution of individual initial conditions is impossible, the best we can hope for is a statistical theory which predicts the likelihood of the variable x_n taking on any particular value. In this case the "order in chaos" which is apparent in Figure 3 plays an important role in delineating the range of validity and the structure of statistical descriptions. For example, in applying this analysis to the evolution of biological populations, we see that for conditions corresponding to $a \sim 4$ the population will fluctuate in an apparently random fashion over the entire range but is most likely to lie at either the maximum or minimum values.

And the standard map

Our second example of a nonlinear dynamical system which exhibits a transition from regular to chaotic behavior is the standard map (*14*), described by a pair of nonlinear difference equations

$$x_{n+1} = x_n + y_{n+1} \quad (2)$$

$$y_{n+1} = y_n + k \sin x_n \quad (3)$$

which map the values of the two variables x_n and y_n at time-step n into x_{n+1} and y_{n+1} at time-step $n + 1$. In this case the parameter k in equation 3 controls the magnitude of the nonlinearity.

This map can be used to describe a large number of physical systems. It provides, for example, an approximate description of the one-dimensional motion of a charged particle perturbed by a broad spectrum of oscillating fields, where x_n and y_n denote the position and velocity of the particle at a discrete time $t = n$ and k is a measure of the electric field amplitude. It also arises naturally as an approximate description of general one-dimensional, nonlinear oscillators subject to periodic perturbations (hence the name "standard map").

As the nonlinear parameter, k, is increased, the evolution of this map exhibits, like the logistic map, a dramatic transformation from regular, predictable motion to chaotic, statistical behavior. As a consequence, detailed numerical and analytical investigations of this classical mechanical system have played, and continue to play, an important role in studies of the microscopic foundations of classical statistical mechanics.

The simplest physical system described by this pair of coupled, nonlinear difference equations is a rigid rotor, such as the one depicted in Figure 4, which is subject to sudden kicks at regular time intervals. In this case the variable x_n corresponds to the angle of the rotor and y_n to the angular velocity immediately after the nth kick, and equations 2 and 3 are just Newton's equations for this classical mechanical system. The kick can be either forward or backward, depending on the sign of $\sin x_n$, and the maximum strength of the kick is determined by the size of the nonlinear parameter, k.

Equations 2 and 3 provide an exact, deterministic description of the evolution of the "phase-point" (x_n, y_n) in the two-dimensional x–y "phase-space" which is uniquely determined by the initial condition (x_0, y_0). For example, if we set $k = 0$ and look at the motion of the rotor at stroboscopic intervals of time, then the angular velocity would remain constant at $y_n = y_0$ and the angle x_n would increase by y_0 each unit of time. A graph of the point (x_n, y_n) in the x–y phase-space of this dynamical system would show a sequence of dots lying in a straight, horizontal line of constant y_n. The first graph in Figure 5 shows a computer-generated "phase-space portrait" (also known as a Poincaré section) for several values of y_0 with $k = 0$. (For convenience we have taken advantage of the natural periodicity of the angle x to restrict the range of x to the interval $[0, 2\pi]$ by evaluating equation 2 modulo 2π.) In fact, an analytical solution which describes this regular behavior for the linear difference equations (linear when $k = 0$) can easily be determined. However, for nonzero k the standard map is no longer integrable and does not admit closed-form analytical solutions for x_n and y_n at an arbitrary time $t = n$. In these cases we must rely heavily on intuition derived from numerical "experiments" to develop new methods of analysis.

We can exploit several symmetries which significantly reduce the complexity of the analysis. The first symmetry is the fact that the map is naturally periodic in y with period 2π. (If we increment y by 2π on both sides of equation 3, its value remains unchanged.) We have already noted that x is an angle variable which is only defined modulo 2π. Therefore, for the purposes of graphic analysis it is convenient to evaluate both equations 2 and 3 modulo 2π so that the evolution of x_n and y_n is restricted to a square in the x–y phase-space with sides of length 2π.

The graphs in Figure 5 show phase-space portraits

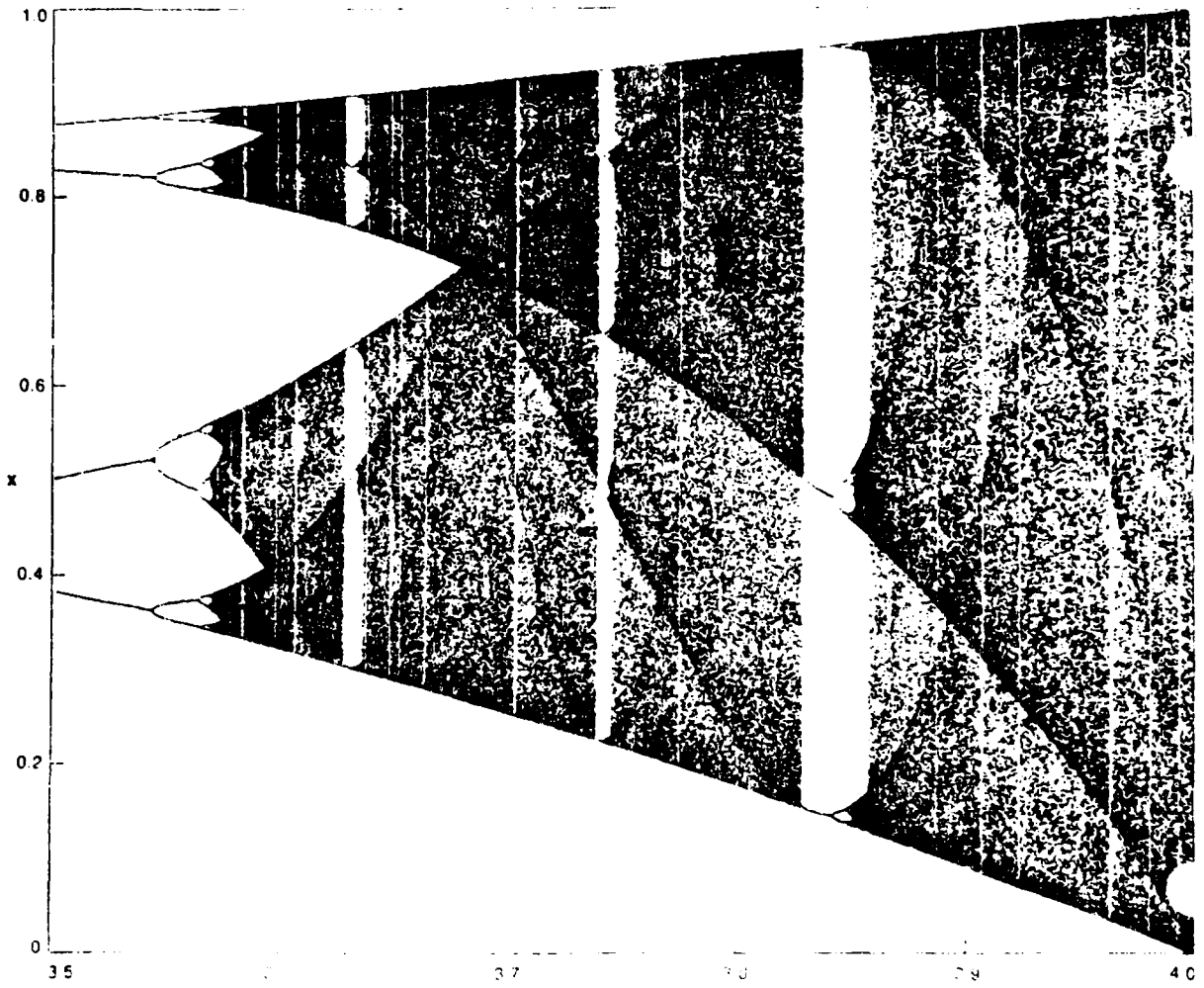


Figure 3. A paradigm in the field of nonlinear dynamics, this bifurcation diagram shows the long-time behavior of the logistic map for values of a between 3.5 and 4. The graph is generated by numerically iterating the map for different values of a and plotting several hundred successive values of x after initial conditions have died out. The result is a graphic display of the underlying structure of chaos—patterns that show an orderly progression from regular to chaotic behavior in any system that can be modeled by equation 1 Where the long-time evolution of the map converges to a periodic cycle of period N, the diagram shows N discrete values for x (at $a = 3.5$, for example, there are four discrete values for x; the system eventually settles down to a period-four oscillation among those four values). Where the evolution is chaotic, the values of x cover continuous intervals, and the darkness of the shading represents the relative probability that x will visit a particular region.

of the restricted dynamics for increasing values of the nonlinear parameter, k. This series of graphs clearly shows the transition from regular to irregular chaotic behavior as the strength of the kicks is increased. (The speckled region in the graph for $k = 2$ is covered by a single orbit.) These graphs of the restricted motion are extremely useful, since initial conditions that show regular or chaotic behavior in the restricted phase space will also exhibit regular and chaotic behavior, respectively, in the unrestricted dynamics.

A second important property which simplifies the mathematical analysis is that the standard map (like most equations describing nondissipative, classical mechanical systems) is [illegible], and the evolution of x_n and y_n preserves area [illegible]

easy to [illegible] think of the map as a coordinate transformation from variables (x_n, y_n) to variables (x_{n+1}, y_{n+1}). From elementary calculus we know that under a change of variables infinitesimal areas and volumes stretch or contract by the magnitude of the Jacobian determinant J of the coordinate transformation. A straightforward evaluation of J for the standard map shows that it is equal to 1. This property permits the application of a wide variety of mathematical methods and theorems on Hamiltonian systems which have been developed over the [illegible] hundred years in the study of celestial mechanics [5].

[illegible] k, the long-term dynamics would [illegible] attracting set in phase space with zero [illegible] a point or [illegible] complex

manifest in [illegible] a "strange attractor" (6, 16). In our simple example of area-added dissipation (friction) to the mapping equations 3 by

$$y_{n+1} = \lambda y_n + k \sin x_n \qquad (4)$$

with $\lambda < 1$ decreasing the velocity each time-step due to friction, then in the absence of any kicks, $k = 0$, every initial condition would evolve to the attracting set defined by the line $y = 0$. However, for large enough values of k the kicks can overcome the friction, and the attracting set can be much more complicated. For example, the upper diagram in Figure 6 shows the outline of the "strange attractor" for $\lambda = 0.1$ and $k = 8.8$ (17). The

We could always imagine that in principle, by stirring very carefully, we can separate the cream from the coffee after it has been thoroughly mixed

reason this attracting set is considered to be strange is that if we magnify a section of any strand of the attractor, we find that it is composed of many strands which in turn are composed of many strands, ad infinitum. The second diagram in Figure 6 is a magnification of a section of this attractor showing this "self-similar" structure on a smaller scale.

The structure's dimension (which is not necessarily an integer), of these fascinating "fractal" attractors are the subject of much current research in nonlinear dynamics. This research has many possible applications, including the description of chaotic behavior in dissipative systems such as turbulent flows, chemical oscillators, or neural networks (1). The interested reader should refer to the excellent review article by Ed Ott (16) and the beautiful book by Benoit Mandelbrot (18).

Returning to the nondissipative standard map, we note that in the absence of an attractor, the phase-point (x_n, y_n) can in principle wander anywhere in the available phase-space. However, we have seen that when $k = 0$, the evolution of the phase-point is confined to a horizontal line. For nonzero k the angular velocity is perturbed by kicks and ceases to be a constant of motion. We might then expect the phase-point to explore all of phase-space. Nevertheless, the phase-space portraits in Figure 5 clearly show that this is not necessarily the case.

This "experimental" result is further substantiated by a remarkable theorem for Hamiltonian systems known as the Kolmogorov, Arnold, Moser (or KAM) theorem (15, 19). This theorem states that if you take an integrable Hamiltonian system (such as the standard map with $k = 0$) and add a nonintegrable perturbation ($k \neq 0$), then sufficiently small perturbations approximate constants of motion will survive and the evolution of the dynamical system will remain regular (if somewhat distorted) for most initial conditions. Although the [illegible] proof of this theorem requires that the perturbation be extremely small, Dr. Percival has described its magnitude as comparable to the "gravitational [illegible] in Australia." You can see

that in Figure 5 the phase-space orbits remain quite regular for fairly large values of k and that the perturbations must become rather large before the evolution of a single phase-trajectory begins to fill large regions of phase-space, as it does in the graph for $k = 2$.

In practice this transition from mostly regular behavior to global chaos as k is increased has tremendous physical significance. For example, numerical experiments show that for small values of k the angular velocity and kinetic energy of the kicked rotor may increase and decrease but remain confined to a restricted range of values for all time. However, for large values of k the velocity and energy can wander over all of phase-space. If in this case we remove the restriction to velocities on the interval $[0, 2\pi]$, we find that the rotor's velocity and energy can wander to arbitrarily large values. Despite the fact that there are no "random" forces at play, this diffusion in energy appears for all intents and purposes to be a random walk. Since the standard map also provides a model for the interaction of charged particles with a broad spectrum of oscillating electrical fields, this deterministic diffusion in energy provides an important means of heating high-temperature, low-density fusion plasmas where "random" particle collisions are too rare to mediate in the irreversible transfer of energy from the fields to the particles (19).

The numerical experiments indicate that this transition from confined to diffusive motion occurs for $k_c \sim 1$. This observation has led to the development of a series of approximate theories of ever increasing sophistication and accuracy for predicting the critical perturbation strength for the onset of global chaos in general nonlinear systems. At present the best theoretical prediction (20) for the standard map is $k_c \leq 63/64 = 0.984375$, which is very close to the best numerical estimate (21) of $k_c \sim 0.971635406$.

The chains of elliptical "island" structures which proliferate at $k \sim 1$ (Fig. 5) play a very important role in determining this transition to global stochasticity. These regular structures in the nonlinear dynamics result from resonances between the motion of the nonlinear oscilla-

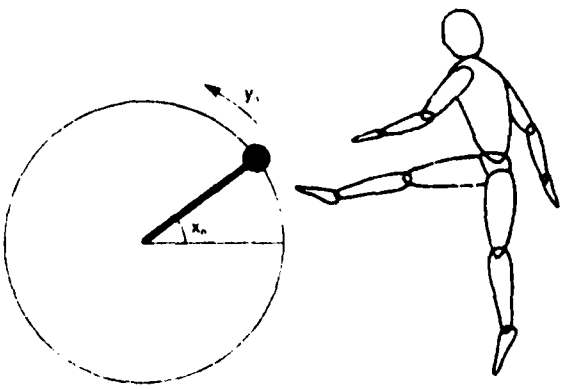


Figure 4. The simplest physical system described by the standard map (equations 2 and 3) is a rigid rotor subject to periodic kicks. The term x_n represents the angular position of the rotor at the time of the nth kick; y_n represents the angular velocity of the rotor just before the nth kick. The strength and direction of the kicks are determined by the nonlinear term $k \sin x_n$.

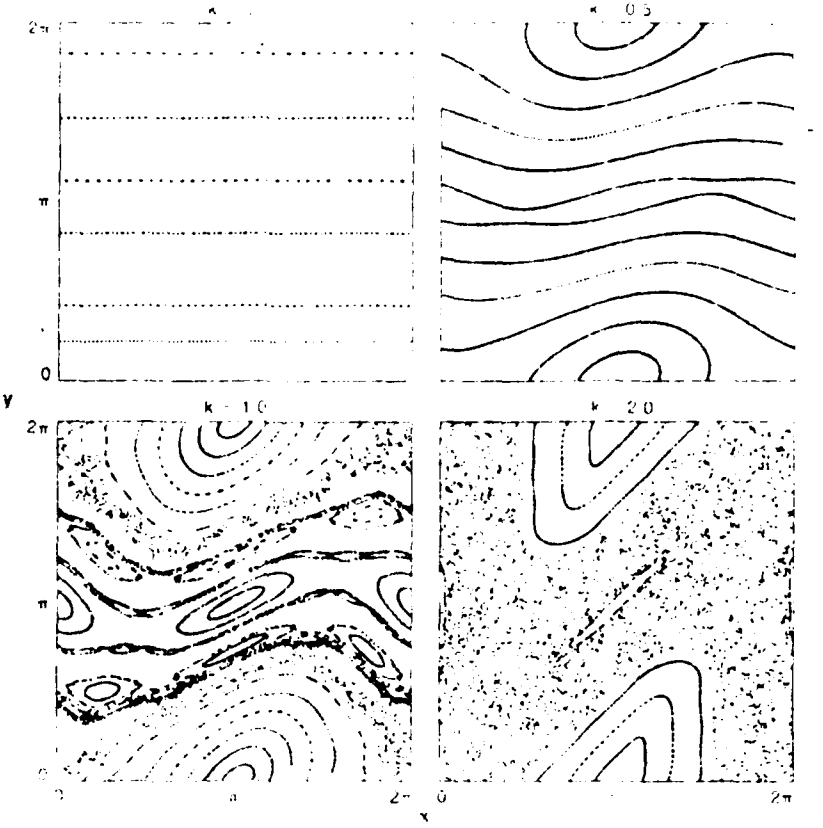


Figure 5. Phase-space portraits (Poincaré sections) for the standard map, shown here for four different values of the nonlinear parameter k, are analogous to the bifurcation diagram for the logistic map (Fig. 3), in that they make it easier to see transitions from regular to chaotic evolution. These figures are generated by numerically iterating the standard map for several different initial conditions (x_0, y_0) and plotting several hundred of the succeeding points (x_n, y_n) in the x–y phase-space. The graph at the upper left shows the regular, integrable dynamics for $k = 0$ (which corresponds, in Fig. 4, to a kick of zero strength, so that the angular velocity of the rotor is constant). When k is increased to 0.5, the trajectories for various initial conditions are still regular and nearly integrable, as guaranteed by the Kolmogorov-Arnold-Moser theorem for small values of k. A mixture of chaotic and regular trajectories appears when $k = 1$. The graph for $k = 2$ is dominated by chaotic evolution: a single trajectory can wander over large regions of phase-space, although some islands of stability persist.

tor (the rotor) and [illegible] turbation. At the center [illegible] land" is a phase-point [illegible] recurs after q iterations of [illegible] where q also happens to be the number of islands in a chain surrounding the distance from $x = 0$ to 2π. The reader can readily verify that the point $(\pi, 0)$, which lies at the center of the large island [illegible] period [illegible] for the [illegible] elliptical island can be found at the top of the y graph.) It is also easy to check that the phase-point (π, π), which lies at the center of the smaller two-island chain across the center of the graph for $k = 1$, is a periodic orbit of period 2; the intermediate values of the rotor angle and velocity correspond to the point $(2\pi, \pi)$, which is the same as $(0, \pi)$ because of the periodicity of the map.

The islands surrounding these periodic orbits correspond to nearby orbits which are trapped in nonlinear resonances. Since those trapped orbits will also oscillate within the trapping region, the period of perturbation will also generate island structures within these regular regions and these islands in turn will generate smaller island chains, *ad infinitum*. For $k \leq 1$ these higher order resonances are extremely narrow and only a few can be discerned at the resolution of Figure 5. However, using Hamiltonian perturbation theory we find that the individual island chains increase in width as $k^{q/2}$, so we would expect catastrophic consequences when k exceeds 1. In fact, the disorder which occurs in the case of large numbers of resonances interacting is [illegible] global chaos (22).

For $k > 1$ the approximate constants of motion are destroyed for most initial conditions, and the corresponding phase-space trajectories are no longer confined to smooth curves but can wander [illegible]

regions of phase-space (like the orbit for $k = 2$ in Fig. 5). In the next section I will show that these orbits exhibit the same local instability and extreme sensitivity to initial conditions as the irregular orbits of the logistic map and that this "random" dynamics meet the conditions required by the definition of chaos. Unfortunately, few rigorous mathematical results are available at present for moderately realistic physical models like the standard map; however, since the map can be easily iterated for many millions of time-steps, the numerical evidence can be very convincing. In fact, one numerical study reported the results of a calculation with as many as 10^{12} iterations of the standard map (23).

One of the difficulties faced by a rigorous mathematical analysis is that phase-space is often divided into both regular regions (inside resonant island structures) as well as chaotic regions for most realistic systems. In particular, the standard map already exhibits bands of chaotic orbits for very small values of k, although the KAM theory guarantees that they are very narrow. These increase in size as k increases until k exceeds k_c, after which the chaotic regions expand until they consume most of phase-space. For example, the bands of chaos are too narrow to be seen in Figure 5 when $k = 0.5$ but begin to appear at $k = 1$ and dominate the phase-space at $k = 2$. Moreover, periodic orbits with stable island structures may persist in the chaotic regime. For example, Figure 5 shows that an island of stability persists around the fixed point at $(\pi, 0)$ for $k = 2$; however, it [illegible] away by the chaotic sea when [illegible]

175

[illegible] March-Apr [illegible]

20

Chaos

The graphs of the angular dynamics generated by the logistic and standard maps provide a picture book of chaos. Like many nonlinear systems in nature, these mathematical models exhibit behavior which appears to be random despite the fact that the equations of motion are fully deterministic. But if the motions are fully determined and the systems are relatively simple, where does this complex behavior come from? What are the symptoms that allow us to identify chaos when we see it? And what are the real differences, if any, between such deterministic chaotic behavior and random processes? To describe more clearly this disease called classical chaos, we must delve a little deeper into the mathematical theory of dynamical systems.

We have already defined what we mean by deterministic behavior in dynamical systems, namely, their evolution is completely determined by the initial conditions and the equations of motion prescribed by the laws of physics. But what do we mean by random behavior? Our intuitive notion of a random or chance process, such as the roll of a die, the flip of a coin, or the spin of a roulette wheel, is a process which exhibits irregular behavior that is not determined by any laws and defies prediction (24). However, this concept would not be very useful if it were not for the fact that statistical properties of such systems, such as the average behavior over many trials after many repetitions, are very well described by the calculus of probabilities and the so-called laws of chance (24). Therefore, the traditional definition of an idealized random or, more precisely, stochastic process is a dynamical system which can be described only in terms of average properties determined by an appropriate probability distribution.

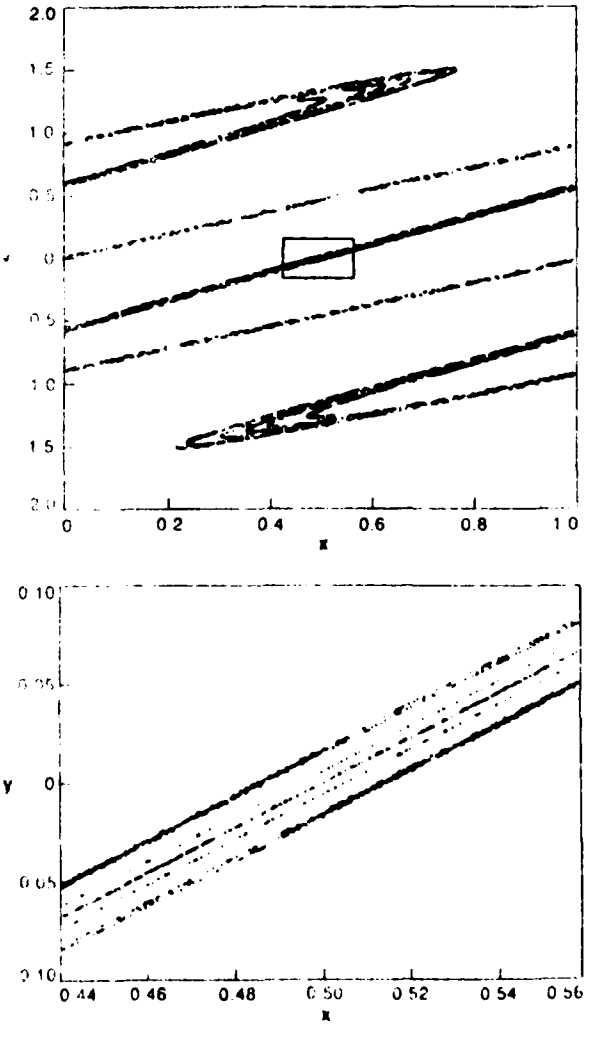
Statistical methods based on the calculus of probabilities and the mathematical theory of stochastic processes have been successfully applied to a wide variety of physical problems. The most spectacular example is the theory of classical statistical mechanics developed by Maxwell, Boltzmann, Gibbs, and Einstein. It provides the physical foundation for the theory of thermodynamics, which accurately describes much of the macroscopic world. However, we now have two antithetical descriptions of the evolution of molecules in gases or atoms in a solid: one deterministic and the other probabilistic; and we must somehow reconcile the foundations of statistical mechanics, which is the connection between them.

[illegible] conditions the long-term evolution of a [illegible] dynamical system can converge to an attracting set in phase [illegible] which is called, in two or more dimensions, a strange attractor. For example, if a dissipation function is added to the rotor depicted in Figure [illegible] that equation 1 replaces equation 3, the diagram at the top shows the strange attractor for the damped, kicked rotor when $\kappa = 0.1$ and $k = 8.8$. The long-term evolution of [illegible] for these parameters. The attractor has a fractal "dust" structure, which is maintained on all scales and which gives this object a noninteger, "fractal," dimension. The [illegible] at the bottom magnifies the region enclosed by the [illegible] and reveals [illegible] the self-similar structure of the central strand of this [illegible].

[illegible] gets to [illegible] probabilistic laws of statistical mechanics and [illegible] deterministic laws of classical [illegible] have both led to a new branch of mathematics, called ergodic theory, which provides a means of classifying different deterministic dynamical systems with irregular behavior (19, 25, 26). In particular, this classification scheme defines symptoms for a hierarchy of different classes of random-like behavior, "statistical diseases," of increasing severity.

Dynamical systems with the mildest disease are called ergodic (25, 26). These are systems that come near almost every possible state over time but do so in a regular manner. For example, the evolution of the standard map for $k = 0$ is completely described by equation 2, since the angular velocity, y_n, is a constant of motion. If the initial angular velocity, y_0, is an irrational multiple of 2π, then the angle variable, x_n, will eventually cover the entire interval $[0, 2\pi]$ in an ordered and predictable way. This system is merely ergodic.

Although there has been considerable confusion in the physical literature, ergodicity alone is not sufficient to justify the application of the probabilistic methods of statistical mechanics, since ergodicity alone does not assure that nonequilibrium distributions evolve toward



equilibrium (25, 26). However, dynamical systems with a more severe disease, the so-called Kolmogorov systems or K-systems, are irregular enough to rigorously justify a statistical description (26, 27).

K-systems exhibit the mathematical property known as "mixing" with "positive Kolmogorov-Sinai entropy." The "mixing" behavior is a precise characterization of what you observe when you stir cream in your coffee, although many nonlinear dynamicists prefer the example of rum and Coke (28). "Positive Kolmogorov-Sinai entropy" is an essential technical condition which is difficult to verify directly for a given dynamical system. However, in practice this means that the dynamical system exhibits extreme sensitivity to initial conditions, so that two trajectories started at nearby initial conditions diverge at an exponential rate. This rate is measured by the "average Liapunov exponent," which is equivalent to the Kolmogorov-Sinai entropy and can be easily computed (29, 30). Because of this extreme sensitivity to initial conditions, the evolution of deterministic K-systems defies long-time prediction (like the weather), since small errors or uncertainties in the initial conditions give rise to time-evolutions which are completely different.

We can now define chaos as the behavior of deterministic dynamical systems which exhibit these symptoms of mixing behavior with a positive Kolmogorov-Sinai entropy or, equivalently, a positive average Liapunov exponent.

For example, for one-dimensional maps (like the logistic map) of the form $x_{n+1} = f(x_n)$, the average Liapunov exponent is defined to be

$$\lambda = \lim_{N \to \infty} \frac{1}{N} \sum_{n=0}^{N-1} \ln \left| \frac{df}{dx}(x_n) \right| \qquad (5)$$

Figure 7 shows a graph of λ versus a for the same range of the nonlinear parameter a as displayed in the bifurcation diagram of Figure 3. Here we clearly see a correspondence between periodic orbits and positive values of the average Liapunov exponent and between periodic orbits and sharp dips in λ.

In particular for $a = 4$ the average Liapunov exponent can be calculated exactly by taking advantage of a remarkable coordinate transformation. If we define a new variable

$$y_n = (2/\pi) \sin^{-1}(\sqrt{x_n}) \qquad (6)$$

then the logistic map equation 1, transforms to the "tent map"

$$y_{n+1} = \begin{cases} 2y_n & 0 \le y_n \le 0.5 \\ 2(1 - y_n) & 0.5 \le y_n \le 1 \end{cases} \qquad (7)$$

Here we see that $\ln|dF/dx| = \ln 2$ for all x so that $\lambda = \ln 2 \sim 0.693 > 0$. Since the Kolmogorov-Sinai entropy is invariant under coordinate changes (26), this proves that the logistic map with $a = 4$ is a K-system and thereby meets our definition of a chaotic dynamical system. It can also be rigorously shown that the logistic map is a K-system and therefore chaotic for many values of $a > a_\infty = 3.57\ldots$, which is consistent with the numerical evidence displayed in Figure 7.

The average Liapunov exponent can also be calculated for dynamical systems in higher dimensions, like the standard map, although the algorithm is more complicated than that for one-dimensional maps (30).

The root of the disease lies in the mathematical pathologies of the real numbers

For example, a computation of the average Liapunov exponent for the standard map shows that for orbits in the regular regions of the phase-space of Figure 5, $\lambda \sim 0$; in the irregular regions, $\lambda > 0$. Unfortunately, very few realistic systems have been rigorously proved to be K-systems. Consequently, the justification for classifying much irregular behavior as chaos depends on the accumulation of numerical evidence and on experience with a few idealized mathematical models which are known to be K-systems.

Using this technical definition of chaos, we now see that chaotic dynamical systems can exhibit many of the attributes of idealized random systems; namely, their evolution is unpredictable because of their extreme sensitivity to initial conditions, and their average properties can be described using statistical methods. However, when we observe irregular phenomena in nature, such as turbulent flow in fluids, we don't always perform averages over time or over an ensemble of initial conditions. Rather, we often observe a single realization of the dynamical process evolving from a specific (though imprecisely known) initial condition which nevertheless

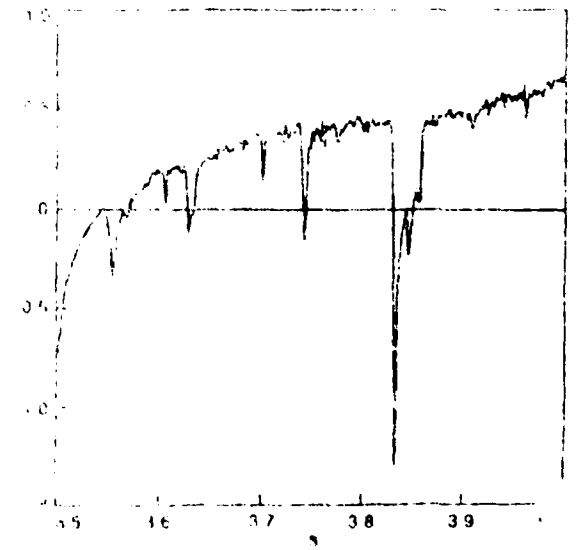


Figure 7. The average Liapunov exponent (λ, equation 5) defines in precise mathematical terms a system's sensitivity to initial conditions: when λ is positive, small changes in initial conditions lead to large divergences in the long-term evolution. The average Liapunov exponent for the logistic map is numerically computed and plotted here for the same values of a shown in Figure 3. This graph verifies that chaotic behavior in Figure 3, the bifurcation diagram, corresponds to positive values for λ, whereas regular (periodic) [illegible]

177

22

appears as random as a sequence of coin flips. Is it possible that deterministic but chaotic dynamical systems can also account for the random appearance of individual realizations of these physical systems? The answer is yes. Using the definition of a random sequence provided by algorithmic complexity theory (7), we will see that the evolution of a chaotic dynamical system can be indistinguishable from a sequence of coin flips and that these completely determined systems can be as irregular as any idealized random system. (This latter conclusion begs the question of whether any idealized random systems exist in the world of classical physics and whether the apparent randomness we observe and exploit in statistical theories is just the chaotic behavior of some underlying deterministic dynamical system.)

Algorithmic complexity theory defines the complexity, K_N, of a sequence of N numbers as the length of the shortest computer program that can generate the sequence (7, 31, 32). This length is conveniently measured

Under chaotic conditions the use of pesticides, price controls, or arms control agreements will not necessarily yield the desired outcomes

in terms of the number of bits of information required to input the program, which is proportional to the number of lines of FORTRAN (or any other programming language) plus the number of bits required to specify any numerical inputs or parameters in the program, such as the number of elements in the sequence, N. In particular, the minimum program size required to generate a sequence of numbers of length N is at least $\log_2 N$, since this is the number of bits required to specify the length of the sequence in binary notation. Moreover, if we consider binary sequences of 0s and 1s so that an output sequence with N elements corresponds to N bits of information, then the maximum value for K_N is of the order of N, since the computer program can simply read the N-bit sequence as input and then output the same sequence. (The programming commands to copy the input add only a constant contribution to K_N, which varies for different computers but is negligibly small in the limit of large N.)

A random sequence is defined to be a sequence with maximal complexity, $K_N \sim N$. A nonrandom sequence can be generated by a shorter program which takes advantage of any order or regularity in the sequence. For example, a sequence consisting of all 1s, corresponding to a sequence of coin flips where heads appears every time, can be generated by the computer program "Print 1, N times," which can be programmed with $\sim \log_2 N$ bits of information. However, a sequence of 1s and 0s with no apparent order, which is most efficiently generated by simply making a copy of it, has maximal complexity, $K_N \sim N$.

This definition of a random sequence, which arose in the work of Kolmogorov, Chaitin, and Solmonov in information theory (7, 31, 32), is in complete agreement with our intuitive conception of a random sequence. Certainly a sequence with any apparent order, such as consecutive 1s, would not be considered to be very random, whereas a sequence that has no regular patterns and can be specified only by a program of length $\sim N$ is likely to meet our intuitive criteria for randomness. In fact, for infinite sequences Martin-Löf has proved that these random sequences will satisfy every conceivable statistical test for randomness (32).

What then, is the complexity of the time-sequences generated by chaotic dynamical systems? Consider for convenience the one-dimensional map on the unit interval

$$x_{n+1} = 2x_n \quad \text{Mod } 1 \tag{8}$$

which is closely related to the tent map and consequently to the logistic map via the coordinate-transformation equation 6. Using equation 5, the average Liapunov exponent for equation 8 is easily determined to be $\ln 2 > 0$; so this map is a K-system and therefore chaotic. Now, if we examine the action of this chaotic dynamical system on initial conditions represented in binary, $x_0 = 0.101001110100111\ldots$, then the multiplication of x_0 by 2 just shifts the "binary" point to the right and the Mod 1 throws away any integer part of x to the left of the "binary" point. Therefore, successive iterations of this "register shift" simply read off successive binary digits in the initial condition. In particular, if we call "heads" when the value of $x_n > \frac{1}{2}$ (i.e., the leading digit is 1) and "tails" when $x_n < \frac{1}{2}$ (in which case the leading digit is 0), then the evolution of the map will generate, from every initial condition, a sequence of heads and tails which resembles the tossing of a coin. But when will these sequences appear random? The answer is again provided by Martin-Löf, who also proved that almost all initial conditions on the unit interval have a random binary-digit sequence (32). Therefore, the deterministic shift map will almost always generate a random sequence which is indistinguishable from the outcome of an idealized coin toss. Moreover, the same conclusions can be generalized to the tent map (and equivalently the logistic map) and all other chaotic dynamical systems. (Of course if you try to implement equation 8 on a digital computer, only short sequences can be studied, because the shift map quickly runs into the precision limits of the computer, which represents initial conditions with only $\sim$ 30 or 60 binary digits in single and double precision, respectively.)

Is physics conquering chaos, or chaos undermining physics?

The definitions and examples in the previous sections show that nonlinear dynamical systems can exhibit all the attributes of an idealized random process. Moreover, the theory of algorithmic complexity reveals that the origins of chaotic behavior in nonlinear dynamical systems and perhaps in nature itself lie in the randomness of almost all real numbers.

In other words, chaotic dynamical systems are mathematical models which "read" initial conditions. They are like the compulsive librarians in Borges's Library of Babel (where books containing every possible combination of letters are shelved), who read every word and character in the books under their care, whereas regular or nonchaotic systems are like the

Figure 8. On the New York Stock Exchange, the use of computers has decreased the frenzied shouting on the floor while increasing the volume of trading—but prices seem more volatile than ever. Even if the prices of stocks are completely determined by initial conditions—that is, if the system is mechanistic—the behavior of the market on a given day might still satisfy the mathematical definition of chaos. There would be no faster way to compute the outcome than to watch the market itself perform on that day. (Photograph © Four By Five.)

casual readers, who just read the titles and skim the text (33). The unpredictability of chaotic dynamical systems arises from the fact that slight errors or changes in the initial conditions correspond to different books in the library which tell different stories.

More generally, if nonlinear models describing the evolution of biological populations, economic prices, armament stockpiles, or turbulent flows in fluids can exhibit chaotic behavior, then we may be incapable, in practice, of predicting the behavior of these systems or their response to external forces, since any errors or perturbations will grow exponentially. For example, under chaotic conditions the use of pesticides, price controls, or arms control agreements will not necessarily yield the desired outcomes (fig. 8).

Another manifestation of the unpredictability of chaotic dynamical systems is that the time evolution is computationally irreducible (34). There is no faster way of finding out how a chaotic system will evolve than to watch its evolution. The dynamical system itself is its own fastest computer. Of all the books in the Library of Babel the only way you can [illegible] the contents is to read the entire book to the end. (Unfortunately, most, in fact almost all, of the books appear to be gibberish and make very uninteresting reading. However, somewhere in the library is a collection which contains the complete [illegible] of the universe.)

Chaotic dynamical systems are also like football games. Even with the largest imaginable digital computer you could not predict the outcome with certainty. The players themselves provide the fastest analog computation of the evolution of this dynamical system. Because of the complexity and unpredictability of chaos, direct numerical simulations of football games and turbulent flows are likely to remain unpractical with even the largest supercomputers. However, we can nevertheless compute reliable odds or probabilities for the outcomes of these processes. As a consequence, probabilistic and statistical theories provide a natural description of average properties of chaotic dynamical systems. (An entertaining account of how several well-known chaos theorists used their knowledge of nonlinear dynamics to improve their odds at roulette is provided in ref. 35.)

One of the most surprising properties of chaotic dynamical systems is that these deterministic models are often very simple. The realization that complex behavior does not require complex mathematical models is one of the most significant contributions of nonlinear dynamics. Since simple models can yield complex, irregular behavior, we can actually hope to develop theoretical descriptions of a wide variety of apparently random, unpredictable natural phenomena using mathematical models which exhibit deterministic chaos. However, although recent progress has suggested partial solutions to the nineteenth-century problems of the origin of turbulence in fluids and the microscopic foundations of statistical mechanics, many old problems remain, and some new and very profound questions have been raised.

For example, among the old problems, the discovery of chaos has not miraculously solved the problem of turbulence in fluids. But we now have new methods of

There is no faster way of finding out how a chaotic system will evolve than to watch its evolution. The system itself is its own fastest computer.

characterizing turbulent behavior, such as the measurement of the average Liapunov exponent or the fractal dimension of the strange attractor associated with turbulence, and we have a much better understanding of why the theoretical and numerical description of the evolution of turbulent flows is so difficult (36).

Moreover, although chaos explains how average properties of nonlinear dynamical systems can exhibit an irreversible approach to thermodynamic equilibrium, it does not account for why individual systems in nature always appear to exhibit the irreversible evolution mandated by the second law of thermodynamics. Since the equations of motion of classical mechanics are deterministic and invariant under time-reversal, we could always imagine that, in principle, by stirring very carefully, we can return the cream from the coffee after it has been

thoroughly mixed. All that nonlinear dynamics tells us is that this reversal of time-evolution will be extremely difficult, since any errors or uncertainties will guarantee failure. The missing ingredient required for a complete justification of the foundations of classical statistical mechanics is an argument for why such errors are inevitable. (Certainly, any numerical simulation of the evolution and reversal of a chaotic dynamical system will fail to recover the initial state, because whenever the machine rounds off a number it automatically introduces a slight change in the system which gives a completely different and unpredictable result.) A provocative discussion of the relationship of chaos to the second law of thermodynamics can be found in Prigogine and Stengers's *Order out of Chaos* (*37*).

What are the real differences, if any, between deterministic chaotic behavior and random processes?

While confronting these remaining problems, nonlinear dynamics has also identified some new, uniquely twentieth-century problems which may account for some of these failures. The first problem is that, since chaotic dynamical systems essentially read initial conditions, they are exquisitely sensitive to the infinities and infinitesimals manifest in the continuum of real numbers which underlie almost all mathematical descriptions of natural phenomena. In contrast, regular systems, such as those studied in almost every textbook, are relatively insensitive to the mathematical pathologies of infinitely long digit-strings.

The difficulty with the continuum of real numbers lies in the fact that, although most real numbers can be proved to have random digit-strings, it is impossible to prove that a given digit-string is random. You simply can never exhaust all the possible tests for underlying order. This is a specific example of a class of true statements which cannot be proved, statements first shown to exist by Gödel in his celebrated incompleteness theorem (*38*). For a clear discussion of the connection between random digit-strings and Gödel's incompleteness theorem, see (*39*). Moreover, by definition these numbers cannot be computed by any algorithm shorter than the digit string itself. As a consequence, most real numbers are uncomputable. Therefore, now that our understanding of chaotic dynamical systems has revealed that the root of the discrepancies in these mathematical pathologies of the real numbers, Joe Ford has suggested that these uncomputable and undefinable objects should be excised from any meaningful physical theory (*7*). In addition to providing some logical consistency in the description of natural phenomena, this restriction might also provide the missing argument for the validity of the second law of thermodynamics. For example, if we assume that nature is a finite-state computer (or Turing machine), then the computable truncation of real numbers would provide the coarse-graining necessary to ensure the irreversibility of chaotic systems.

Such [illegible] would [illegible] in natural sciences, and a number of research groups have already begun to explore the possibilities of so-called cellular automata models for natural phenomena which are defined on discrete sets of numbers (*39*). However, it is possible that the scale at which the truncation of real numbers occurs may be so small that no practical consequences of the distinction between continuum and discrete theories can be deduced or verified. In that case the issue of the ultimate discretization of the real world will pass from the domain of physics to that of philosophy. Nevertheless, cellular automata are a fascinating subject in their own right and promise to play an important role in future studies of nonlinear dynamical systems.

The second fundamental question which arises from our improved understanding of classical nonlinear systems is whether chaos persists in microscopic physical systems, such as atoms and molecules, where the theory of quantum mechanics is expected to apply (*40, 41*). The difficulty here lies in the fact that the Schrödinger equation for the evolution of the quantum mechanical wave function is a linear equation which, strictly speaking, is incapable of exhibiting the chaotic behavior of nonlinear classical systems. Since quantum mechanics is presumed to be the fundamental theory for all physical systems, and since the predictions of quantum theory must agree with those of classical mechanics at the limit of the highest quantum numbers, according to Bohr's correspondence principle, one of these physical descriptions—classical chaos or quantum mechanics—threatens to undermine the other. Does this mean that the role of classical chaos in explaining the origins of turbulence and the foundations of statistical mechanics is merely an illusion? That Bohr's correspondence principle is invalid for systems that are classically chaotic? And that there isn't any problem with the continuum of real numbers after all?

The answers to these questions are naturally the goals of much current research. Preliminary results indicate that the evolution of the quantum mechanical wave function appears to mimic chaotic behavior for very long times (*42, 43*), in many cases longer than the age of the universe (*44*). Nevertheless, without chaos we have lost some of the necessary ingredients for the foundations of statistical mechanics. The validity of the correspondence principle, which guided the early development of quantum mechanics, also remains an outstanding problem, although recent experiments on the ionization of highly excited hydrogen atoms exposed to intense electromagnetic radiation (which study the behavior of a quantum system that is classically chaotic) suggest that the correspondence principle is remarkably robust (*45, 46*).

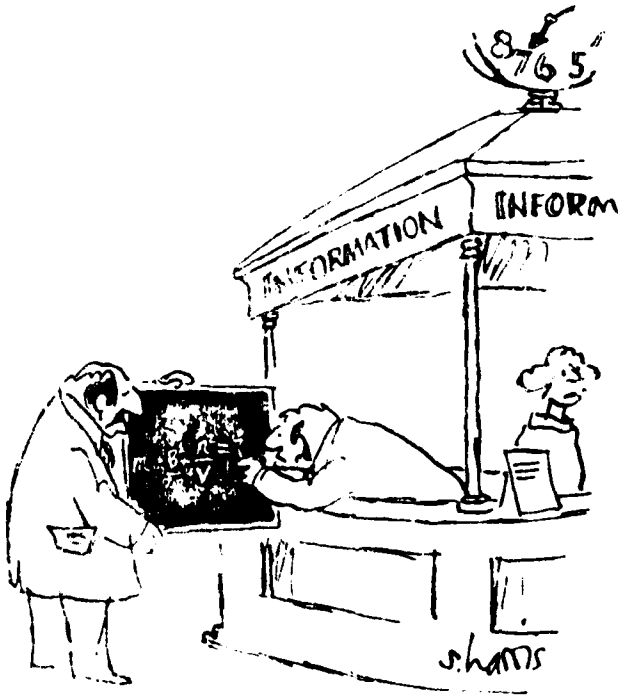
In conclusion, we have seen how deceptively simple mathematical models for nonlinear dynamical systems, like the logistic and standard maps, have provided new hope for the description of the complexity and chaos which surround us in the natural world. However, these and more recent studies have also opened a Pandora's box of new problems which ask profound and disturbing questions about the proper mathematical description of both macroscopic and microscopic natural phenomena and which promise to lie at the forefront of scientific research for many years to come.

References

1. See articles in *Order in Chaos, Physica 7D*, North Holland, 1983.
2. R. M. May. 1976. Simple mathematical models with very complicated dynamics. *Nature* 261:459.
3. L. Glass, M. R. Guevara, and A. Shrier. 1983. Bifurcation and chaos in a periodically stimulated cardiac oscillator. *Physica 7D*, p. 89.
4. R. V. Jensen and R. Urban. 1984. Chaotic price behavior in a nonlinear cobweb model. *Econ. Lett.* 15:235.
5. A. M. Saperstein. 1984. Chaos—a model for the outbreak of war. *Nature* 309:303.
6. J. P. Crutchfield, J. D. Farmer, N. H. Packard, and R. S. Shaw. 1986. Chaos. *Sci. Am.* 255, Dec., p. 46.
7. J. Ford. 1983. How random is a coin toss? *Phys. Today* 36, Apr., p. 40.

   ——— 1986. Chaos. Solving the unsolvable, predicting the unpredictable! In *Chaotic Dynamics and Fractals*, Academic Press.

   ——— 1986. What is chaos, that we should be mindful of it? In *The New Physics*, ed. S. Capelin and P. C. W. Davies. Cambridge Univ. Press.
8. M. J. Feigenbaum. 1983. Universal behavior in nonlinear systems. *Physica 7D*, p. 16.
9. P. Cvitanovic. 1984. *Universality in Chaos*. Bristol, UK: Adam Hilger.
10. H. G. Schuster. 1984. *Deterministic Chaos*. Weinheim, FRG: Physik-Verlag.
11. P. Collet and J.-P. Eckmann. 1980. *Iterated Maps on the Interval as Dynamical Systems*. Birkhauser.
12. R. V. Jensen and C. R. May, p. 1985. Images of critical points of nonlinear maps. *Phys. Rev.* [illegible] 1222.
13. C. Grebogi, E. Ott, and J. Yorke. 1983. Crises, sudden changes in chaotic attractors, and transient chaos. *Physica 7D*, p. 181.
14. B. V. Chirikov. 1979. A universal instability of many-dimensional oscillator systems. *Phys. Rep.*, 52:263.
15. V. I. Arnold. 1978. *Mathematical Methods of Classical Mechanics*. Springer-Verlag.
16. E. Ott. 1981. Strange attractors and chaotic motions of dynamical systems. *Rev. Mod. Phys.* 53:[illegible].
17. R. V. Jensen and C. R. [illegible]. Statistical properties of chaotic dynamical systems [illegible] strange attractors. *Physica 4D*, p. 183.
18. B. B. Mandelbrot. 1982. *The Fractal Geometry of Nature*. W. H. Freeman.
19. A. J. Lichtenberg and M. A. Lieberman. 1983. *Regular and Stochastic Processes*. Springer-Verlag.
20. R. S. MacKay and I. C. Percival. 1985. Converse KAM: Theory and practice. *Comm. Math. Phys.* 98:469.
21. J. M. Greene. 1979. A method for determining a stochastic transition. *J. Math. Phys.* 20:1183.
22. G. M. Zaslavskii and B. V. Chirikov. 1972. Stochastic instability of nonlinear oscillations. *Soviet Physics USPEKHI* 14:549.
23. C. F. F. Karney. 1983. Long-time correlations in the stochastic regime. *Physica 8D*, p. 360.
24. H. Poincaré. 1952. *Science and Method*. Dover.
25. J. L. Lebowitz and O. Penrose. 1973. Modern ergodic theory. *Phys. Today*, Feb., p. 23.
26. Ya. G. Sinai. 1977. *Introduction to Ergodic Theory*. Princeton Univ. Press.
27. N. S. Krylov. 1979. *Works on the Foundations of Statistical Physics*. Princeton Univ. Press.
28. V. I. Arnold and A. Avez. 1968. *Ergodic Problems of Classical Mechanics*. Benjamin.
29. Ya. B. Pesin. 1977. Characteristic Lyapunov exponents and smooth ergodic theory. *Russ. Math. Surv.* 32:55.
30. G. Benettin, L. Galgani, and J.-M. Strelcyn. 1976. Kolmogorov entropy and numerical experiments. *Phys. Rev. A* 14:2338.
31. G. J. Chaitin. 1975. Randomness and mathematical proof. *Sci. Am.* 232, May, [illegible].
32. ——— 1982. Godel's theorem and information. *Int. J. Theor. Phys.* 21:941.
33. P. Martin-Lof. 1966. The definition of random sequences. *Info. Contr.* 9:602.
34. J. L. Borges. 1964. The Library of Babel. In *Labyrinths*, p. 51. New Directions.
35. S. Wolfram. 1985. Undecidability and intractability in theoretical physics. *Phys. Rev. Lett.* 54:735.
36. T. A. Bass. 1985. *The Eudemonic Pie: Or Why Would Anyone Play Roulette without a Computer in His Shoe?* Houghton Mifflin.
37. K. Sreenivasan. 1985. Transitions and turbulence in fluid flows and low dimensional chaos. In *Frontiers in Fluid Mechanics*, ed. S. H. Davis and J. L. Lumley, p. 41. Springer-Verlag.
38. I. Prigogine and I. Stengers. 1984. *Order out of Chaos*. Bantam.
39. D. Hofstadter. 1979. *Gödel, Escher, Bach: An Eternal Golden Braid*. Basic Books.
40. See articles in *Cellular Automata, Physica 10D*, North Holland, 1984.
41. See articles in *Chaotic Behavior in Quantum Systems*, ed. G. Casati. Plenum, 1985.
42. M. V. Berry. 1983. Semi-classical mechanics of regular and irregular motion. In *Chaotic Behavior of Deterministic Systems*, ed. G. Iooss, R. H. G. Helleman, and R. Stora, p. 171. North Holland.
43. R. V. Jensen and R. Shankar. 1985. Statistical behavior of deterministic quantum systems with few degrees of freedom. *Phys. Rev. Lett.* 54:1879.
44. G. Casati, B. V. Chirikov, I. Guarneri, and D. L. Shepelansky. 1986. Dynamical stability of quantum "chaotic" motion in a hydrogen atom. *Phys. Rev. Lett.* 56:2437.
45. A. Peres. 1982. Recurrence phenomena in quantum mechanics. *Phys. Rev. Lett.* 49:1118.
46. K. A. H. van Leeuwen et al. 1985. Microwave ionization of hydrogen atoms: Experiment versus classical dynamics. *Phys. Rev. Lett.* 55:2231.
47. R. V. Jensen. In press. Chaos in atomic physics. In *Proceedings of the 14th International Conference on Atomic Physics (ICAP 14)*, ed. H. Narumi. North Holland.



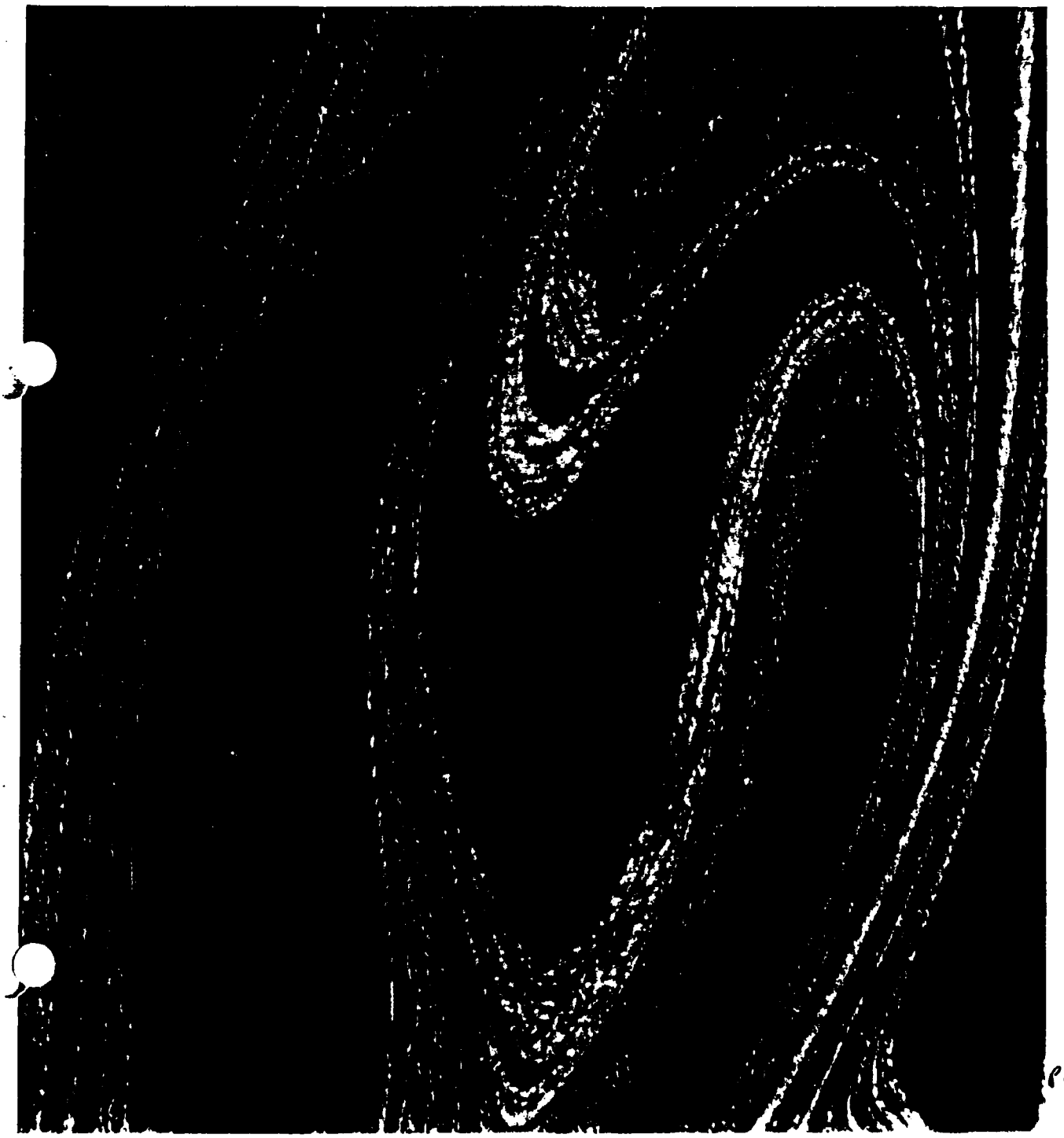
Appendix 4:

"Chaos, Strange Attractors, and Fractal Basin Boundaries in Nonlinear Dynamics", Celso Grebogi, Edward Ott, James A. Yorke, Science, 30 Oct 1987, Vol. 238, pages 631-638.

'ERICAN
OCIATION FOR THE
ADVANCEMENT OF
SCIENCE

SCIENCE

30 OCTOBER 1987 $2.50
VOL. 238 ■ PAGES 585–718



P631

Chaos, Strange Attractors, and Fractal Basin Boundaries in Nonlinear Dynamics

CELSO GREBOGI, EDWARD OTT, JAMES A. YORKE

Recently research has shown that many simple nonlinear deterministic systems can behave in an apparently unpredictable and chaotic manner. This realization has broad implications for many fields of science. Basic developments in the field of chaotic dynamics of dissipative systems are reviewed in this article. Topics covered include strange attractors, how chaos comes about with variation of a system parameter, universality, fractal basin boundaries and their effect on predictability, and applications to physical systems.

IN THIS ARTICLE WE PRESENT A REVIEW OF THE FIELD OF chaotic dynamics of dissipative systems including recent developments. The existence of chaotic dynamics has been discussed in the mathematical literature for many decades with important contributions by Poincaré, Birkhoff, Cartwright and Littlewood, Levinson, Smale, and Kolmogorov and his students, among others. Nevertheless, it is only recently that the wide-ranging impact of chaos has been recognized. Consequently, the field is now undergoing explosive growth, and many applications have been made across a broad spectrum of scientific disciplines—ecology, economics, physics, chemistry, engineering, fluid mechanics, to name several. Specific examples of chaotic time dependence include convection of a fluid heated from below, simple models for the yearly variation of insect populations, stirred chemical reactor systems, and the determination of limits on the length of reliable weather forecasting. It is our belief that the number of these applications will continue to grow.

We start with some basic definitions of terms used in the rest of the article.

Dissipative system. In Hamiltonian (conservative) systems such as arise in Newtonian mechanics of particles (without friction), phase space volumes are preserved by the time evolution. (The phase space is the space of variables that specify the state of the system.) Consider, for example, a two-dimensional phase space (q, p), where q denotes a position variable and p a momentum variable. Hamilton's equations of motion take the set of initial conditions at time $t = t_0$ and evolve them in time to the set at time $t = t_1$. Although the shapes of the sets are different, their areas are the same. By a dissipative system we mean one that does not have this property (and cannot be made to have this property by a change of variables). Areas should typically decrease (dissipate) in time so that the area of the final set would be less than the area of the initial set. As a consequence of this, dissipative systems typically are characterized by the presence of attractors.

Attractor. If one considers a system and its phase space, then the initial conditions may be attracted to some subset of the phase space (the attractor) as time $t \to \infty$. For example, for a damped harmonic oscillator (Fig. 1a) the attractor is the point at rest (in this case the origin). For a periodically driven oscillator in its limit cycle the limit set is a closed curve in the phase space (Fig. 1b).

Strange attractor. In the above two examples, the attractors were a point (Fig. 1a), which is a set of dimension zero, and a closed curve (Fig. 1b), which is a set of dimension one. For many other attractors the attracting set can be much more irregular (some would say pathological) and, in fact, can have a dimension that is not an integer. Such sets have been called "fractal" and, when they are attractors, they are called strange attractors. [For a more precise definition see (*1*).] The existence of a strange attractor in a physically interesting model was first demonstrated by Lorenz (*2*).

Dimension. There are many definitions of the dimension d (*3*). The simplest is called the box-counting or capacity dimension and is defined as follows:

$$d = \lim_{\epsilon \to 0} \frac{\ln N(\epsilon)}{\ln(1/\epsilon)} \qquad (1)$$

where we imagine the attracting set in the phase space to be covered by small D-dimensional cubes of edge length ϵ, with D the dimension of the phase space. $N(\epsilon)$ is the minimum number of such cubes needed to cover the set. For example, for a point attractor (Fig. 1a), $N(\epsilon) = 1$ independent of ϵ, and Eq. 1 yields $d = 0$ (as it should). For a limit cycle attractor, as in Fig. 1b, we have that $N(\epsilon) \sim \ell/\epsilon$, where ℓ is the length of the closed curve in the figure (dotted line); hence, for this case, $d = 1$, by Eq. 1. A less trivial example is illustrated in Fig. 2, in the form of a Cantor set. This set is

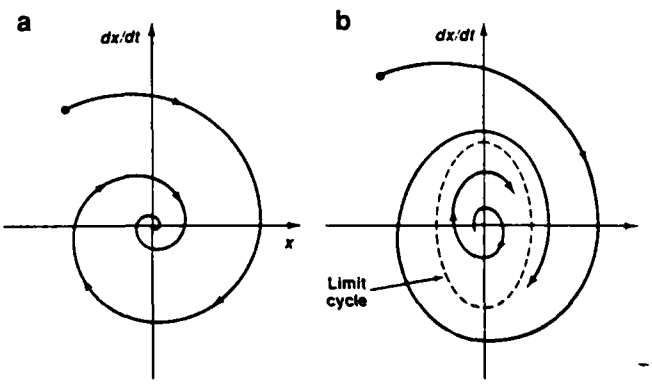


Fig. 1. (**a**) Phase-space diagram for a damped harmonic oscillator. (**b**) Phase-space diagram for a system that is approaching a limit cycle.

C. Grebogi is a research scientist at the Laboratory for Plasma and Fusion Energy Studies, E. Ott is a professor in the departments of electrical engineering and physics, and J. A. Yorke is a professor of mathematics and is the director of the Institute for Physical Science and Technology, University of Maryland, College Park, MD 20742.

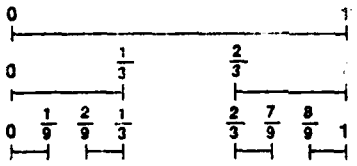


Fig. 2. Construction of a Cantor set.

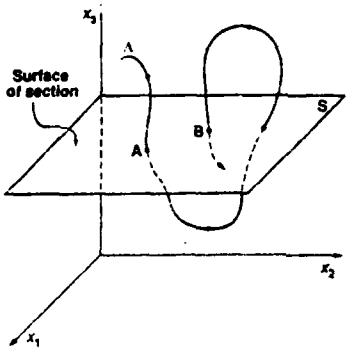


Fig. 3. Poincaré surface of section.

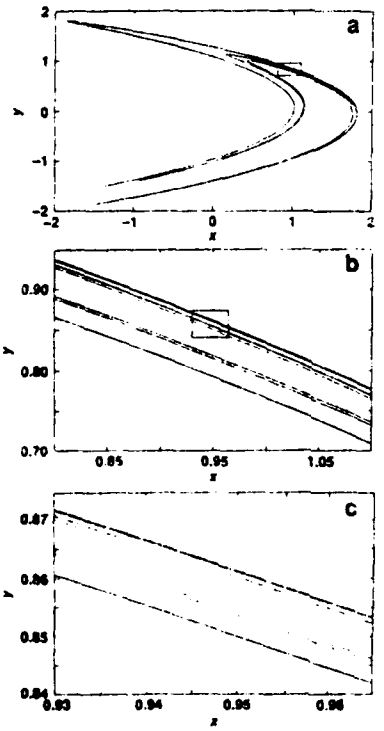


Fig. 4. The Hénon chaotic attractor. (**a**) Full set. (**b**) Enlargement of region defined by the rectangle in (a). (**c**) Enlargement of region defined by the rectangle in (b).

formed by taking the line interval from 0 to 1, dividing it in thirds, then discarding the middle third, then dividing the two remaining thirds into thirds and discarding their middle thirds, and so on ad infinitum. The Cantor set is the closed set of points that are left in the limit of this repeated process. If we take $\epsilon = 3^{-n}$ with n an integer, then we see that $N(\epsilon) = 2^n$ and Eq. 1 (in which $\epsilon \to 0$ corresponds to $n \to \infty$) yields $d = (\ln 2)/(\ln 3)$, a number between 0 and 1, hence, a fractal. The topic of the dimension of strange attractors is a large subject on which much research has been done. One of the most interesting aspects concerning dimension arises from the fact that the distribution of points on a chaotic attractor can be nonuniform in a very singular way. In particular, there can be an arbitrarily fine scaled interwoven structure of regions where orbit trajectories are dense and sparse. Such attractors have been called *multifractals* and can be characterized by subsidiary quantities that essentially give the dimensions of the dense and sparse regions of the attractor. In this review we shall not attempt to survey this work. Several papers provide an introduction to recent work on the dimension of chaotic attractors (*3–5*).

Chaotic attractor. By this term we mean that if we take two typical points on the attractor that are separated from each other by a small distance $\Delta(0)$ at $t = 0$, then for increasing t they move apart exponentially fast. That is, in some average sense $\Delta(t) \sim \Delta(0)\exp(ht)$ with $h > 0$ (where h is called the Lyapunov exponent). Thus a small uncertainty in the initial state of the system rapidly leads to inability to forecast its future. [It is not surprising, therefore, that the pioneering work of Lorenz (*2*) was in the context of meteorology.] It is typically the case that strange attractors are also chaotic [although this is not always so; see (*1, 6*)].

Dynamical system. This is a system of equations that allows one, in principle, to predict the future given the past. One example is a system of first-order ordinary differential equations in time, $d\mathbf{x}(t)/dt = \mathbf{G}(\mathbf{x},t)$, where $\mathbf{x}(t)$ is a D-dimensional vector and $\mathbf{G}$ is a D-dimensional vector function of $\mathbf{x}$ and t. Another example is a map.

Map. A map is an equation of the form $\mathbf{x}_{t+1} = \mathbf{F}(\mathbf{x}_t)$, where the "time" t is discrete and integer valued. Thus, given $\mathbf{x}_0$, the map gives $\mathbf{x}_1$. Given $\mathbf{x}_1$, the map gives $\mathbf{x}_2$, and so on. Maps can arise in continuous time physical systems in the form of a Poincaré surface of section. Figure 3 illustrates this. The plane x_3 = constant is the surface of section (S in the figure), and A denotes a trajectory of the system. Every time A pierces S going downward (as at points A and B in the figure), we record the coordinates (x_1,x_2). Clearly the coordinates of A uniquely determine those of B. Thus there exists a map, B = F(A), and this map (if we knew it) could be iterated to find all subsequent piercings of S.

Chaotic Attractors

As an example of a strange attractor consider the map first studied by Hénon (*7*):

$$x_{n+1} = \alpha - x_n^2 + \beta y_n \tag{2}$$

$$y_{n+1} = x_n \tag{3}$$

Figure 4a shows the result of plotting 10^4 successive points obtained by iterating Eqs. 2 and 3 with parameters $\alpha = 1.4$ and $\beta = 0.3$ (and the initial transient is deleted). The result is essentially a picture of the chaotic attractor. Figure 4, b and c, shows successive enlargements of the small square in the preceding figure. Scale invariant, Cantor set–like structure transverse to the linear structure is evident. This suggests that we may regard the attractor in Fig. 4c, for example, as being essentially a Cantor set of approximately straight parallel lines. In fact, the dimension d in Eq. 1 can be estimated numerically (δ) to be $d \cong 1.26$ so that the attractor is strange.

As another example consider a forced damped pendulum described by the equation

$$d^2\theta/dt^2 + \nu d\theta/dt + \omega_0^2\sin\theta = f\cos(\omega t) \tag{4}$$

where θ is the angle between the pendulum arm and the rest position, ν is the coefficient of friction, ω_0 is the frequency of natural oscillation, and f is the strength of the forcing. In Eq. 4, the first term represents the inertia of the pendulum, the second term represents friction at the pivot, the third represents the gravitational force, and the right side represents an external sinusoidally varying torque of strength f and frequency ω applied to the pendulum at the pivot. In Fig. 5a, we plot the Poincaré surface of section of a strange

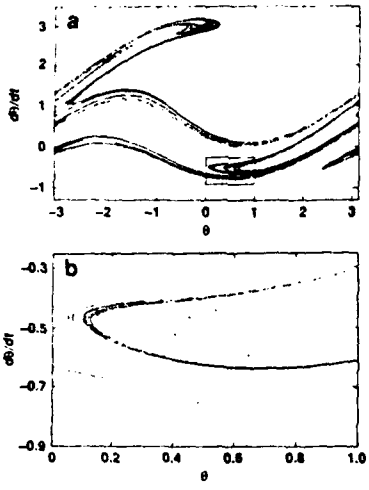


Fig. 5. (**a**) Poincaré surface of section of a pendulum strange attractor. (**b**) Enlargement of region defined by rectangle in (**a**).

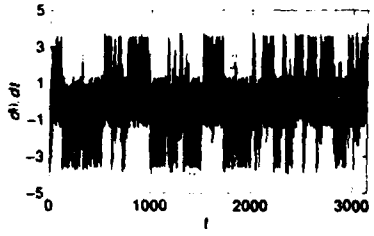


Fig. 6. Chaotic time series for pendulum shown as a plot of angular velocity versus time.

attractor for the pendulum, where we choose $\nu = 0.22$, $\omega_0 = 1.0$, $\omega = 1.0$, and $f = 2.7$ in Eq. 4. This surface of section is obtained by plotting 50,000 dots, one dot for every cycle of the forcing term, that is, one dot at every time $t = t_n = 2\pi n$ (where n is an integer). The strange attractor shown in Fig. 5a exhibits a Cantor set–like structure transverse to the linear structure. This is evident in Fig. 5b, which shows an enlargement of the square region in Fig. 5a. The dimension of this strange attractor in the surface of section is $d = 1.38$. Figure 6 shows the angular velocity $d\theta/dt$ as a function of t for the parameters of Fig. 5. Note the apparently erratic nature of this plot.

In general, the form of chaotic attractors varies greatly from system to system and even within the same system. This is indicated by the sequence of chaotic attractors shown in Fig. 7. All of these attractors were generated from the same map (*9*),

$$\psi_{n+1} = [\psi_n + \omega_1 + \epsilon P_1(\psi_n,\theta_n)] \bmod 1 \quad (5)$$

$$\theta_{n+1} = [\theta_n + \omega_2 + \epsilon P_2(\psi_n,\theta_n)] \bmod 1 \quad (6)$$

where P_1 and P_2 are periodic with period one in both their arguments. The P_1 and P_2 are the same in all of the cases shown in Fig. 7; only the parameters ω_1, ω_2, and ϵ have been varied. The results show the great variety of form and structure possible in chaotic attractors as well as their aesthetic appeal. Since ψ and θ may be regarded as angles, Eqs. 5 and 6 are a map on a two-dimensional toroidal surface. [This map is used in (*9*) to study the transition from quasiperiodicity to chaos.]

Because of the exponential divergence of nearby orbits on chaotic attractors, there is a question as to how much of the structure in these pictures of chaotic attractors (Figs. 4, 5, and 7) is an artifact due to chaos-amplified roundoff error. Although a numerical trajectory will diverge rapidly from the true trajectory with the same initial point, it has been demonstrated rigorously (*10*) in important cases [including the Hénon map (*11*)] that there exists a true trajectory with a slightly different initial point that stays near the noisy trajectory for a long time. [For example, for the Hénon map for a typical numerical trajectory computed with 14-digit precision there exists a true trajectory that stays within 10^{-7} of the numerical trajectory for 10^7 iterates (*11*).] Thus we believe that the apparently fractal structure seen in pictures such as Figs. 4, 5, and 7 is real.

The Evolution of Chaotic Attractors

In dissipative dynamics it is common to find that for some value of a system parameter only a nonchaotic attracting orbit (a limit cycle, for example) occurs, whereas at some other value of the parameter a chaotic attractor occurs. It is therefore natural to ask how the one comes about from the other as the system parameter is varied continuously. This is a fundamental question that has elicited a great deal of attention (*9*, *12*–*19*).

To understand the nature of this question and some of the possible answers to it, we consider Fig. 8a, the so-called bifurcation diagram for the map.

$$x_{n+1} = C - x_n^2 \quad (7)$$

where C is a constant. Figure 8a can be constructed as follows: take $C = -0.4$, set $x_0 = -0.5$, iterate the map 100 times (to eliminate transients), then plot the next 1000 values of x; increase C by a small amount, say 0.001, and repeat what was done for $C = -0.4$; increase again, and repeat; and so on, until $C = 2.1$ is reached. We see from Fig. 8a that below a certain value, $C = C_0 = -0.25$, there is no attractor in $-2 < x < 2$. In fact, in this case all orbits go to $x \to -\infty$, hence the absence of points on the plot. This is also true for C above the "crisis value" $C_c = 2.0$. Between these two values there is an attractor. As C is increased we have an attracting orbit of "period one," which, at $C = 0.75$, bifurcates to a period-two attracting orbit ($x_\alpha \to x_\beta \to x_\alpha \to x_\beta \to \cdots$), which then bifurcates (at $C = 1.25$) to a period-four orbit ($x_a \to x_b \to x_c \to x_d \to x_a \to x_b \to x_c \to x_d \to x_a \to \cdots$). In fact, there are an infinite number of such bifurcations of period 2^n to period 2^{n+1} orbits, and these accumulate as $n \to \infty$ at a finite value of C, which we denote C_∞ (from Fig. 8a, $C_\infty \cong 1.4$). [The practical importance of this phenomenology was emphasized early on by May (*12*).]

What is the situation for $C_\infty < C < C_c$? Numerically what one sees is that for many C values in this range the orbits appear to be chaotic, whereas for others there are periodic orbits. For example, Fig. 8b shows an enlargement of Fig. 8a for C in the range $1.72 < C < 1.82$. We see what appear to be chaotic orbits below $C = C_0^{(3)} = 1.75$. However, just above this value, a period-three orbit appears, supplanting the chaos. The period-three orbit then goes through a period-doubling cascade, becomes chaotic, widens into a three-piece chaotic attractor, and then finally at $C = C_c^{(3)} \cong 1.79$ widens back into a single chaotic band. We call the region $C_0^{(3)} < C < C_c^{(3)}$ a period-three window. (Such windows, but of higher period, appear throughout the region $C_\infty < C < C_c$, but are not as discernible in Fig. 8a because they are much narrower than the period-three window.)

An infinite period-doubling cascade is one way that a chaotic attractor can come about from a nonchaotic one (*13*). There are also two other possible routes to chaos exemplified in Fig. 8, a and b. These are the intermittency route (*14*) and the crisis route (*15*).

Intermittency. Consider Fig. 8b. For C just above $C_0^{(3)}$ there is a period-three orbit. For C just below $C_0^{(3)}$ there appears to be chaotic orbit. To understand the character of this transition it is useful to examine the chaotic orbit for C just below $C_0^{(3)}$. The character of this orbit is as follows. The orbit appears to be a period-three orbit for long stretches of time after which there is a short

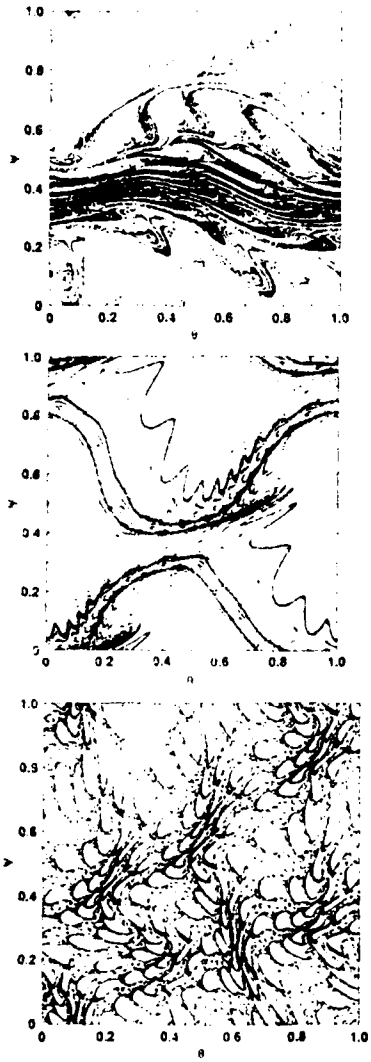


Fig. 7. Sequence of chaotic attractors for system represented by Eqs. 5 and 6. Plot shows iterated mapping on a torus for different values of ω_1, ω_2, and ϵ. (**Top**) $\omega_1 = 0.54657$, $\omega_2 = 0.36736$, and $\epsilon = 0.75$. (**Center**) $\omega_1 = 0.45922$, $\omega_2 = 0.53968$, and $\epsilon = 0.50$. (**Bottom**) $\omega_1 = 0.41500$, $\omega_2 = 0.73500$, and $\epsilon = 0.60$.

burst (the "intermittent burst") of chaotic-like behavior, followed by another long stretch of almost period-three behavior, followed by a chaotic burst, and so on. As C approaches $C_0^{(3)}$ from below, the average duration of the long stretches between the intermittent bursts becomes longer and longer (*14*), approaching infinity and proportional to $(C_0^{(3)} - C)^{-1/2}$ as $C \to C_0^{(3)}$. Thus the pure period-three orbit appears at $C = C_0^{(3)}$. Alternatively we may say that the attracting periodic attractor of period three is converted to a chaotic attractor as the parameter C decreases through the critical value $C_0^{(3)}$. It should be emphasized that, although our illustration of the transition to chaos by way of intermittency is within the context of the period-three window of the quadratic map given by Eq. 7, this phenomenon (as well as period-doubling cascades and crises) is very general, in other systems it occurs for other periods (period one, for example) in easily observable form.

Crises. From Fig. 8a we see that there is a chaotic attractor for $C < C_c = 2$, but no chaotic attractor for $C > C_c$. Thus, as C is lowered through C_c, a chaotic attractor is born. How does this occur? Note that at $C = C_c$ the chaotic orbit occupies the interval $-2 \le x \le 2$. If C is just slightly larger than C_c, an orbit with initial condition in the interval $-2 < x < 2$ will typically follow a chaotic-like path for a long time, after which it finds its way out of the interval $-2 \le x \le 2$, and then rapidly begins to move to large negative x values (that is, it begins to approach $x = -\infty$). This is called a chaotic transient (*15*). The length of a chaotic transient will depend on the particular initial condition chosen. One can define a mean transient duration by averaging over, for example, a uniform distribution of initial conditions in the interval $-2 < x < 2$. For the quadratic map, this average duration is

$$\tau \sim 1/(C - C_c)^{\gamma} \tag{8}$$

with the exponent γ given by $\gamma = 1/2$. Thus as C approaches C_c from above, the lifetime of a chaotic transient goes to infinity and the transient is converted to a chaotic attractor for $C < C_c$. Again, this type of phenomenon occurs widely in chaotic systems. For example, the model of Lorenz (*2*) for the nonlinear evolution of the Rayleigh-Bénard instability of a fluid subjected to gravity and heated from below has a chaotic onset of the crisis type and an accompanying chaotic transient. In that case, γ in Eq. 8 is $\gamma \sim 4$ (*20*). In addition, a theory for determining the exponent γ for two-dimensional maps and systems such as the forced damped pendulum has recently been published (*21*). Thus we have seen that the period doubling, intermittency, and crisis routes to chaos are illustrated by the simple quadratic map (Eq. 7).

We emphasize that, although a map was used for illustrating these routes, all of these phenomena are present in continuous-time systems and have been observed in experiments. As an example of chaotic transitions in a continuous time system, we consider the set of three autonomous ordinary differential equations studied by Lorenz (*2*) as a model of the Rayleigh-Bénard instability,

$$dx/dt = Py - Px \tag{9}$$

$$dy/dt = -xz + rx - y \tag{10}$$

$$dz/dt = xy - bz \tag{11}$$

where P and b are adjustable parameters. Fixing $P = 10$ and $b = 8/3$ and varying the remaining parameter, r, we obtain numerical solutions that are clear examples of the intermittency and crisis types of chaotic transitions discussed above. We illustrate these in Fig. 9, a through d; the behavior of this system is as follows:

1) For r between 166.0 and 166.2 there is an intermittency transition from a periodic attractor ($r = 166.0$, Fig. 9a) to a chaotic attractor ($r = 166.2$, Fig. 9b) with intermittent turbulent bursts. Between the bursts there are long stretches of time for which the orbit oscillates in nearly the same way as for the periodic attractor (*14*) (Fig. 9a).

2) For a range of r values below $r = 24.06$ there are two periodic attractors, that represent clockwise and counterclockwise convections. For r slightly above 24.06, however, there are three attractors, one that is chaotic (shown in the phase space trajectory in Fig. 9c), whereas the other two attractors are the previously mentioned periodic attractors. The chaotic attractor comes into existence as r increases through $r = 24.06$ by conversion of a chaotic transient. Figure 9d shows an orbit in phase space executing a chaotic transient before settling down to its final resting place at one of the periodic attractors. Note the similarity of the chaotic transient trajectory in Fig. 9d with the chaotic trajectory in Fig. 9c.

The various routes to chaos have also received exhaustive experimental support. For instance, period-doubling cascades have been observed in the Rayleigh-Bénard convection (*22, 23*), in nonlinear circuits (*24*), and in lasers (*25*); intermittency has been observed in the Rayleigh-Bénard convection (*26*) and in the Belousov-Zhabotinsky reaction (*27*); and crises have been observed in nonlinear circuits (*28–30*), in the Josephson junction (*31*), and in lasers (*32*).

Finally, we note that period doubling, intermittency, and crises do not exhaust the possible list of routes to chaos. (Indeed, the

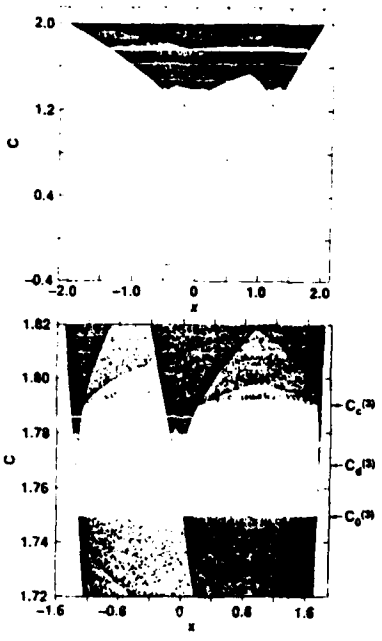


Fig. 8. (**Top**) Bifurcation diagram for the quadratic map. (**Bottom**) Period-three window for the quadratic map.

routes are not all known.) In particular, chaotic onsets involving quasiperiodicity have not been discussed here (*9*, *16*, *18*).

Universality

Universality refers to the fact that systems behave in certain quantitative ways that depend not on the detailed physics or model description but rather only on some general properties of the system. Universality has been examined by renormalization group (*33*) techniques developed for the study of critical phenomena in condensed matter physics. In the context of dynamics, Feigenbaum (*13*) was the first to apply these ideas, and he has extensively developed them, particularly for period doubling for dissipative systems. [See (*17*) for a collection of papers on universality in nonlinear dynamics.] For period doubling in dissipative systems, results have been obtained on the scaling behavior of power spectra for time series of the dynamical process (*34*), on the effect of noise on period doubling (*35*), and on the dependence of the Lyapunov exponent (*36*) on a system parameter. Applications of the renormalization group have also been made to intermittency (*19*, *37*), and the breakdown of quasiperiodicity in dissipative (*18*) and conservative (*38*) systems.

As examples, two "universal" results can be stated within the context of the bifurcation diagrams (Fig. 8, a and b). Let C_n denote the value of C at which a period 2^n cycle period doubles to become a period 2^{n+1} cycle. Then, for the bifurcation diagram in Fig. 8a, one obtains

$$\lim_{n\to\infty} \frac{C_n - C_{n-1}}{C_{n+1} - C_n} = 4.669201\ldots \qquad (12)$$

The result given in Eq. 12 is not restricted to the quadratic map. In fact, it applies to a broad class of systems that undergo period doubling cascades (*13*, *39*). In practice such cascades are very common, and the associated universal numbers are observed to be well approximated by means of fairly low order bifurcations (for example, $n = 2,3,4$). This scaling behavior has been observed in many experiments, including ones on fluids, nonlinear circuits, laser systems, and so forth. Although universality arguments do not explain why cascades must exist, such explanations are available from bifurcation theory (*40*).

Figure 8b shows the period-three window within the chaotic range of the quadratic map. As already mentioned, there are an infinite number of such periodic windows. [In fact, they are generally believed to be dense in the chaotic range. For example, if k is prime, there are $(2^k - 2)/(2k)$ period-k windows.] Let $C_0^{(k)}$ and $C_c^{(k)}$ denote the upper and lower values of C bounding the period-k window and let $C_d^{(k)}$ denote the value of C at which the period-k attractor bifurcates to period $2k$. Then we have that, for typical k windows (*41*),

$$\lim_{k\to\infty} \frac{C_c^{(k)} - C_0^{(k)}}{C_d^{(k)} - C_0^{(k)}} \to 9/4 \qquad (13)$$

In fact, even for the $k = 3$ window (Fig. 8b) the 9/4 value is closely approximated (it is $9/4 - 0.074\ldots$). This result is universal for one-dimensional maps (and possibly more generally for any chaotic dynamical process) with windows.

Fractal Basin Boundaries

In addition to chaotic attractors, there can be sets in phase space on which orbits are chaotic but for which points near the set move away from the set. That is, they are repelled. Nevertheless, such chaotic repellers can still have important macroscopically observable effects, and we consider one such effect (*42*, *43*) in this section.

Typical nonlinear dynamical systems may have more than one time-asymptotic final state (attractor), and it is important to consider the extent to which uncertainty in initial conditions leads to uncertainty in the final state. Consider the simple two-dimensional phase space diagram schematically depicted in Fig. 10. There are two attractors denoted A and B. Initial conditions on one side of the boundary, Σ, eventually asymptotically approach B; those on the other side of Σ eventually go to A. The region to the left or right of Σ is the basin of attraction for attractor A or B, respectively, and Σ is the basin boundary. If the initial conditions are uncertain by an amount ϵ, then for those initial conditions within ϵ of the boundary we cannot say a priori to which attractor the orbit eventually tends.

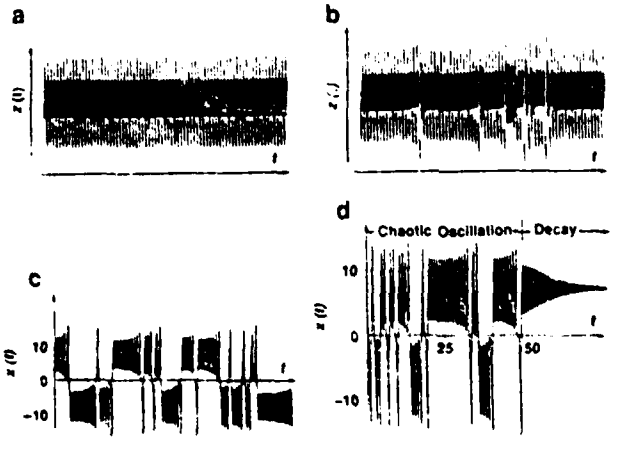


Fig. 9. Intermittency, crisis, and period doubling in continuous time systems. Intermittency in the Lorenz equations (**a**) $r = 166.0$; (**b**) $r = 166.2$. Crisis transition to a chaotic attractor in the Lorenz equations. (**c**) $r = 28$, (**d**) $r = 22$.

Fig. 10. A region of phase space divided by the basin boundary Σ into basins of attraction for the two attractors A and B. Points 1 and 2 are initial conditions with error ϵ.

For example, in Fig. 10, points 1 and 2 are initial conditions with an uncertainty ϵ. The orbit generated by initial condition 1 is attracted to attractor B. Initial condition 2, however, is uncertain in the sense that the orbit generated by 2 may be attracted either to A or B. In particular, consider the fraction of the uncertain phase space volume within the rectangle shown and denote this fraction f. For the case shown in Fig. 10, we clearly have $f \sim \epsilon$. The main point we wish to make in what follows is that, from the point of view of prediction, much worse scalings of f with ϵ frequently occur in nonlinear dynamics. Namely, the fraction can scale as

$$f \sim \epsilon^{\alpha} \tag{14}$$

with the "uncertainty exponent" α satisfying $\alpha < 1$ (*42, 43*). In fact, $\alpha << 1$ is fairly common. In such a case, a substantial reduction in the initial condition uncertainty, ϵ, yields only a relatively small decrease in the uncertainty of the final state as measured by f.

Although α is equal to unity for simple basin boundaries, such as that depicted in Fig. 10, boundaries with noninteger (fractal) dimension also occur. We use here the capacity definition of dimension, Eq. 1. In general, since the basin boundary divides the phase space, its dimension d must satisfy $d \geq D - 1$, where D is the dimension of the phase space. It can be proven that the following relation between the index α and the basin boundary dimension holds (*42, 43*)

$$\alpha = D - d \tag{15}$$

For a simple boundary, such as that depicted in Fig. 10, we have $d = D - 1$, and Eq. 15 then gives $\alpha = 1$, as expected. For a fractal basin boundary, $d > D - 1$, and Eq. 15 gives $\alpha < 1$.

We now illustrate the above with a concrete example. Consider the forced damped pendulum as given by Eq. 4. For parameter values $\nu = 0.2$, $\omega_n = 1.0$, $\omega = 1.0$, and $f = 2.0$, we find numerically that the only attractors in the surface of section (θ, $d\theta/dt$) are the fixed points $(-0.477, -0.609)$ and $(-0.471, 2.037)$. They represent solutions with average counterclockwise and clockwise rotation at the period of the forcing. The cover shows a computer-generated picture of the basins of attraction for the two fixed point attractors. Each initial condition in a 1024 by 1024 point grid is integrated until it is close to one of the two attractors (typically 100 cycles). If an orbit goes to the attractor at $\theta = -0.477$, a blue dot is plotted at the corresponding initial condition. If the orbit goes to the other attractor, a red dot is plotted. Thus the blue and red regions are essentially pictures of the basins of attraction for the two attractors to the accuracy of the grid of the computer plotter. Fine-scale structure in the basins of attraction is evident. This is a consequence of the Cantor-set nature of the basin boundary. In fact, magnifications of the basin boundary show that, as we examine it on a smaller and smaller scale, it continues to have structure.

We now wish to explore the consequences for prediction of this infinitely fine-scaled structure. To do this, consider an initial condition (θ, $d\theta/dt$). What is the effect of a small change ϵ in the θ-coordinate? Thus we integrate the forced pendulum equation with the initial conditions (θ, $d\theta/dt$), (θ, $d\theta/dt + \epsilon$), and (θ, $d\theta/dt - \epsilon$) until they approach one of the attractors. If either or both of the perturbed initial conditions yield orbits that do not approach the same attractor as the unperturbed initial condition, we say that (θ, $d\theta/dt$) is uncertain. Now we randomly choose a large number of initial conditions and let f denote the fraction of these that we find to be uncertain. As a result of these calculations, we find that $f \sim \epsilon^{\alpha}$ where $\alpha = 0.275 \pm 0.005$. If we assume that f, determined in the way stated above, is approximately proportional to f [there is some support for this conjecture from theoretical work (*44*)], then $\alpha = \bar{\alpha}$. Thus, from Eq. 15, the dimension of the basin boundary is $d = 1.725 \pm 0.005$. We conclude, from Eq. 14, that in this case if we are to gain a factor of 2 in the ability to predict the asymptotic final state of the system, it is necessary to increase the accuracy in the measurement of the initial conditions by a factor substantially greater than 2 (namely by $2^{1/0.275} \cong 10$). Hence, fractal basin boundaries ($\alpha < 1$) represent an obstruction to predictability in nonlinear dynamics.

Some representative works on fractal basin boundaries, including applications, are listed in (*42–47*). Notable basic questions that have recently been answered are the following:

1) How does a nonfractal basin boundary become a fractal basin boundary as a parameter of the system is varied (*45*)? This question is similar, in spirit, to the question of how chaotic attractors come about.

2) Can fractal basin boundaries have different dimension values in different regions of the boundary, and what boundary structures lead to this situation? This question is addressed in (*46*) where it is shown that regions of different dimension can be intertwined on an arbitrarily fine scale.

3) What are the effects of a fractal basin boundary when the system is subject to noise? This has been addressed in the Josephson junction experiments of (*31*).

Conclusion

Chaotic nonlinear dynamics is a vigorous, rapidly expanding field. Many important future applications are to be expected in a variety of areas. In addition to its practical aspects, the field also has fundamental implications. According to Laplace, determination of the future depends only on the present state. Chaos adds a basic new aspect to this rule: small errors in our knowledge can grow exponentially with time, thus making the long-term prediction of the future impossible.

Although the field has advanced at a great rate in recent years, there is still a wealth of challenging fundamental questions that have yet to be adequately dealt with. For example, most concepts developed so far have been discovered in what are effectively low-dimensional systems; what undiscovered important phenomena will appear only in higher dimensions? Why are transiently chaotic motions so prevalent in higher dimensions? In what ways is it possible to use the dimension of a chaotic attractor to determine the dimension of the phase space necessary to describe the dynamics? Can renormalization group techniques be extended past the borderline of chaos into the strongly chaotic regime? These are only a few questions. There are many more, and probably the most important questions are those that have not yet been asked.

REFERENCES AND NOTES

1. C. Grebogi, E. Ott, S. Pelikan, J. A. Yorke, *Physica* 13D, 261 (1984).
2. E. N. Lorenz, *J. Atmos. Sci.* 20, 130 (1963).
3. J. D. Farmer, E. Ott, J. A. Yorke, *Physica* 7D, 153 (1983).
4. J. Kaplan and J. A. Yorke, *Lecture Notes in Mathematics No. 730* (Springer Verlag, Berlin, 1978), p. 228; L. S. Young, *Ergodic Theory Dyn. Syst.* 1, 381 (1981).
5. P. Grassberger and I. Procaccia, *Phys. Rev. Lett.* 50, 346 (1983); H. G. E. Hentschel and I. Procaccia, *Physica* 8D, 435 (1983); P. Grassberger, *Phys. Lett.* A97, 227 (1983); T. C. Halsey *et al.*, B. I. Shraiman, *Phys. Rev. A* 33, 1141 (1986); C. Grebogi, E. Ott, J. A. Yorke, *ibid.* 36, 3522 (1987).
6. A. Bondeson *et al.*, *Phys. Rev. Lett.* 55, 2103 (1985); E. J. Romeiras, A. Bondeson, E. Ott, T. M. Antonsen, C. Grebogi, *Physica* 26D, 277 (1987).
7. M. Hénon, *Commun. Math. Phys.* 50, 69 (1976).
8. D. A. Russell, J. D. Hanson, E. Ott, *Phys. Rev. Lett.* 45, 1175 (1980).
9. C. Grebogi, E. Ott, J. A. Yorke, *Physica* 15D, 354 (1985).

10 D. V. Anosov, *Proc. Steklov Ins. Math.* **90** (1967); R. Bowen, *J. Differ. Equations* **18**, 333 (1975).
11. S. M. Hammel, J. A. Yorke, C. Grebogi, *J. Complexity*, **3**, 136 (1987).
12. R. M. May, *Nature (London)* **261**, 459 (1976).
13. M. J. Feigenbaum, *J. Stat. Phys.* **19**, 25 (1978).
14. Y. Pomeau and P. Manneville, *Commun. Math. Phys.* **74**, 189 (1980).
15. C. Grebogi, E. Ott, J. A. Yorke, *Physica* **7D**, 181 (1983).
16. D. Ruelle and F. Takens, *Commun. Math. Phys.* **20**, 167 (1971).
17. P. Cvitanovic, Ed. *Universality in Chaos* (Hilger, Bristol, 1984).
18. For example, S. J. Shenker, *Physica* **5D**, 405 (1982); K. Kaneko, *Progr. Theor. Phys.* **71**, 282 (1984); M. J. Feigenbaum, I. P. Kadanoff, S. J. Shenker, *Physica* **5D**, 370 (1982); D. Rand, S. Ostlund, J. Sethna, E. Siggia, *Physica* **8D**, 303 (1983); S. Kim and S. Ostlund, *Phys. Rev. Lett.* **55**, 1165 (1985); D. K. Umberger, J. D. Farmer, I. I. Satija, *Phys. Lett. A* **114**, 341 (1986); P. Bak, T. Bohr, M. H. Jensen, *Phys. Scr.* **T9**, 50 (1985); P. Bak, *Phys. Today* **39** (No. 12), 38 (1987).
19. I. E. Hirsch, M. Nauenberg, D. J. Scalapino, *Phys. Lett. A* **87**, 391 (1982).
20. J. A. Yorke and E. D. Yorke, *J. Stat. Phys.* **21**, 263 (1979); in *Topics in Applied Physics* (Springer-Verlag, New York, 1981), vol. 45, p. 77.
21. C. Grebogi, E. Ott, J. A. Yorke, *Phys. Rev. Lett.* **57**, 1284 (1986).
22. A. Libchaber and J. Maurer, *J. Phys. (Paris)* **41**, C3-51 (1980); A. Libchaber, C. Laroche, S. Fauve, *J. Phys. (Paris) Lett.* **43**, L-211 (1982).
23. J. P. Gollub, S. V. Benson, J. F. Steinman, *Ann. N.Y. Acad. Sci.* **357**, 22 (1980); M. Giglio, S. Musazzi, U. Perini, *Phys. Rev. Lett.* **47**, 243 (1981).
24. P. S. Linsay, *Phys. Rev. Lett.* **47**, 1349 (1981).
25. F. T. Arecchi, R. Meucci, G. Puccioni, J. Tredicce, *ibid.* **49**, 1217 (1982).
26. M. Dubois, M. A. Rubio, P. Bergé, *ibid.* **51**, 1446 (1983).
27. J. C. Roux, P. DeKepper, H. L. Swinney, *Physica* **7D**, 57 (1983).
28. C. Jeffries and J. Perez, *Phys. Rev. A* **27**, 601 (1983); S. K. Brorson, D. Dewey, P. S. Linsay, *ibid.* **28**, 1201 (1983).
29. H. Ikezi, J. S. deGrasse, T. H. Jensen, *ibid.* **28**, 1207 (1983).
30. R. W. Rollins and E. R. Hunt, *ibid.* **29**, 3327 (1984).
31. M. Iansiti *et al.*, *Phys. Rev. Lett.* **55**, 746 (1985).
32. D. Dangoisse, P. Glorieux, D. Hannequin, *ibid.* **57**, 2657 (1986).
33. K. G. Wilson and J. Kogut, *Phys. Rep.* **C 12**, 75 (1974); B. Hu, *ibid.* **91**, 233 (1982).
34. M. J. Feigenbaum, *Phys. Lett. A* **74**, 375 (1979); R. Brown, C. Grebogi, E. Ott, *Phys. Rev. A* **34**, 2248 (1986); M. Nauenberg and J. Rudnick, *Phys. Rev. B* **24**, 493 (1981); B. A. Huberman and A. B. Zisook, *Phys. Rev. Lett.* **46**, 626 (1981); J. D. Farmer, *ibid.* **47**, 179 (1981).
35. J. Crutchfield, M. Nauenberg, J. Rudnick, *Phys. Rev. Lett.* **46**, 933 (1981); B. Shraiman, C. E. Wayne, P. C. Martin, *ibid.*, p. 935.
36. B. A. Huberman and J. Rudnick, *ibid.* **45**, 154 (1980).
37. B. Hu and J. Rudnick, *ibid.* **48**, 1645 (1982).
38. L. P. Kadanoff, *ibid.* **47**, 1641 (1981); D. F. Escande and F. Doveil, *J. Stat. Phys.* **26**, 257 (1981); R. S. MacKay, *Physica* **7D**, 283 (1983).
39. P. Collet, J. P. Eckmann, O. E. Lanford III, *Commun. Math. Phys.* **76**, 211 (1980).
40. J. A. Yorke and K. A. Alligood, *ibid.* **100**, 1 (1985).
41. J. A. Yorke, C. Grebogi, E. Ott, L. Tedeschini-Lalli, *Phys. Rev. Lett.* **54**, 1095 (1985).
42. C. Grebogi, S. W. McDonald, E. Ott, J. A. Yorke, *Phys. Lett. A* **99**, 415 (1983).
43. S. W. McDonald, C. Grebogi, E. Ott, J. A. Yorke, *Physica* **17D**, 125 (1985).
44. S. Pelikan, *Trans. Am. Math. Soc.* **292**, 695 (1985).
45. C. Grebogi, E. Ott, J. A. Yorke, *Phys. Rev. Lett.* **56**, 1011 (1986); *Physica* **24D**, 243 (1987); F. C. Moon and G.-X. Li, *Phys. Rev. Lett.* **55**, 1439 (1985).
46. C. Grebogi, E. Kostelich, E. Ott, J. A. Yorke, *Phys. Lett.* **A118**, 448 (1986); *Physica* **25D**, 347 (1987); C. Grebogi, E. Ott, J. A. Yorke, H. E. Nusse, *Ann. N.Y. Acad. Sci.* **497**, 117 (1987).
47. C. Mira, *C. R. Acad. Sci.* **288A**, 591 (1979); C. Grebogi, E. Ott, J. A. Yorke, *Phys. Rev. Lett.* **50**, 935 (1983); R. G. Holt and I. B. Schwartz, *Phys. Lett.* **A105**, 327 (1984); I. B. Schwartz, *ibid.* **106**, 339 (1984); I. B. Schwartz, *J. Math. Biol.* **21**, 347 (1985); S. Takesue and K. Kaneko, *Progr. Theor. Phys.* **71**, 35 (1984); O. Decroly and A. Goldbeter, *Phys. Lett.* **A105**, 259 (1984); E. G. Gwinn and R. M. Westervelt, *Phys. Rev. Lett.* **54**, 1613 (1985); *Phys. Rev. A* **33**, 4143 (1986); Y. Yamaguchi and N. Mishima, *Phys. Lett.* **A109**, 196 (1985); M. Napiorkowski, *ibid.* **113**, 111 (1985); F. T. Arecchi, R. Badii, A. Politi, *Phys. Rev. A* **32**, 402 (1985); S. W. McDonald, C. Grebogi, E. Ott, J. A. Yorke, *Phys. Lett.* **A107**, 51 (1985); J. S. Nicolis and I. Tsuda, in *Simulation, Communication, and Control*, S. G. Tzafestas, Ed. (North Holland, Amsterdam, 1985); J. S. Nicolis, *Rep. Prog. Phys.* **49**, 1109 (1986); J. S. Nicolis, *Kybernetes* **14**, 167 (1985).
48. This work was supported by the Air Force Office of Scientific Research, the U.S. Department of Energy, the Defense Advanced Research Projects Agency, and the Office of Naval Research.

COVER Even systems as simple as a periodically forced damped pendulum can have complex behavior. This computer-generated plot shows initial pendulum velocities (measured horizontally) and positions (measured vertically). Orbits starting at points in the red region eventually settle into one type of periodic motion, while orbits starting in the blue region yield a different type of periodic motion. The boundary between these regions is fractal. The lighter the shade of red or blue, the longer it takes to settle into the corresponding motion. See page 632. [Photo courtesy of C. Grebogi, E. Ott, and J. A. Yorke, University of Maryland, College Park, MD 20742]

Appendix 5:

"A Better Way to Compress Images", Michael Barnsley and Alan Sloan, BYTE, January 1988.

Reprinted with permission.

A Better Way to Compress Images

Mathematics is providing a novel technique for achieving compression ratios of 10,000 to 1—and higher

Michael F. Barnsley and Alan D. Sloan

THE NATURAL WORLD is filled with intricate detail. Consider the geometry on the back of your hand: the pores, the fine lines, and the color variations. A camera can capture that detail and, at your leisure, you can study the photo to see things you never noticed before. Can personal computers be made to carry out similar functions of image storage and analysis? If so, then image compression will certainly play a central role.

The reason is that digitized images—images converted into bits for processing by a computer—demand large amounts of computer memory. For example, a high-detail gray-scale aerial photograph might be blown up to a 3½-foot square and then resolved to 300 by 300 pixels per square inch with 8 significant bits per pixel. Digitization at this level requires 130 megabytes of computer memory—too much for personal computers to handle.

For real-world images such as the aerial photo, current compression techniques can achieve ratios of between 2 to 1 and 10 to 1. By these methods, our photo would still require between 65 and 13 megabytes.

In this article, we describe some of the main ideas behind a new method for image compression using fractals. The method has yielded compression ratios in excess of 10,000 to 1 (bringing our aerial photo down to a manageable 13,000 bytes). The color pictures in figures 1 through 5 were encoded using the new technique; actual storage requirements for these images range from 100 to 2000 bytes.

A mathematics research team at the Georgia Institute of Technology is developing the system, with funding provided by the Defense Advanced Research Projects Agency (DARPA) and the Georgia Tech Research Corporation (GTRC). Our description is necessarily simplified, but it will show you how a fractal image-compression scheme operates and how to use it to create exciting images.

Describing Natural Objects

Traditional computer graphics encodes images in terms of simple geometrical shapes: points, line segments, boxes, circles, and so on. More advanced systems use three-dimensional elements, such as spheres and cubes, and add color and shading to the description.

Graphics systems founded on traditional geometry are great for creating pictures of man-made objects, such as bricks, wheels, roads, buildings, and cogs. However, they don't work well at all when the problem is to encode a sunset, a tree, a lump of mud, or the intricate structure of a black spleenwort fern. Think about using a standard graphics system to encode a digitized picture of a cloud: You'd have to tell the computer the address and color attribute of each point in the cloud. But that's exactly what an uncompressed digitized image is—a long list of addresses and attributes.

To escape this difficulty, we need a richer library of geometrical shapes. These shapes need to be flexible and controllable so that they can be made to conform to clouds, mosses, feathers, leaves, and faces, not to mention waving sunflowers and glaring arctic wolves. Fractal geometry provides just such a collection of shapes. For a hint of this, glance at the pictures in *The Fractal Geometry of Nature* by Benoit Mandelbrot, who coined the term *fractal* to describe objects that are very "fractured" (see references for additional books and articles). Some elementary fractal images accompany this article.

Using fractals to simulate landscapes and other natural effects is not new; it has been a primary practical application. For instance, through experimentation, you find that a certain fractal generates a pattern similar to tree bark. Later, when you want to render a tree, you put the tree-bark fractal to work.

What *is* new is the ability to start with an actual image and find the fractals that will imitate it to any desired degree of accuracy. Since our method includes a compact way of representing these fractals, we end up with a highly compressed data set for reconstructing the original image.

Overview of Fractal Compression

We start with a digitized image. Using image-processing techniques such as color separation, edge detection, spectrum analysis, and texture-variation analysis, we break up the image into segments. (Some of the same techniques

continued

Michael F. Barnsley and Alan D. Sloan are professors of mathematics at the Georgia Institute of Technology (Atlanta, GA 30332) and officers of Iterated Systems Inc. (1266 Holly Lane NE, Atlanta, GA 30329).

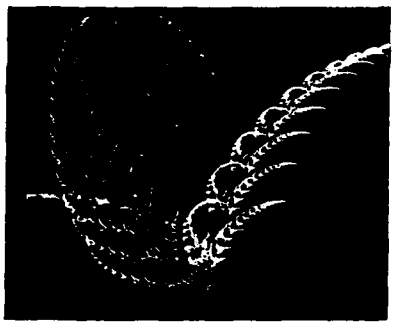


Figure 1: *IFS-encoded color image of three-dimensional ferns (4 transformations, 100 bytes).*

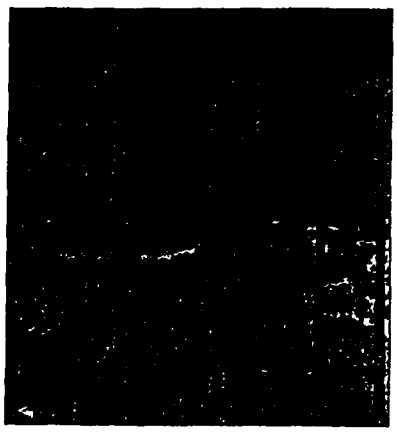


Figure 2: *IFS-encoded color photo of Black Forest, color set adjusted to give winter tones (120 transformations, 2000 bytes).*

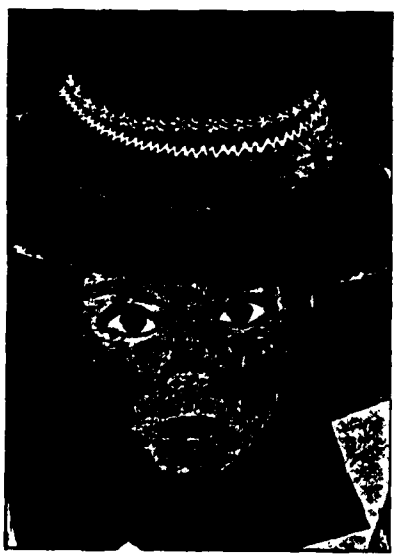


Figure 3: *IFS-encoded color photo of a Bolivian girl (120 transformations, 2000 bytes).*

form the basis for the automatic coloring of black-and-white motion pictures.) A segment might be a fern, a leaf, a cloud, or a fence post. A segment can also be a more complex collection of pixels: A seascape, for example, may include spray, rock, and mist.

We then look up these segments in a library of fractals. The library doesn't contain literal fractals; that would require astronomical amounts of storage. Instead, our library contains relatively compact sets of numbers, called *iterated function system* (IFS) codes, that will reproduce the corresponding fractals. Furthermore, the library's cataloging system is such that images that look alike are close together: Nearby codes correspond to nearby fractals. This makes it feasible to set up automated procedures for searching the library to find fractals that approximate a given target image. A mathematical result known as the Collage Theorem (more on that later) guarantees that we can always find a suitable IFS code—and gives a method for doing so.

Once we have looked up all the segments in our library and found their IFS codes, we can throw away the original digitized image and keep the codes, achieving our compression ratio of 10,000 to 1—or even higher.

Iterated Function Systems

We start by explaining how a set of IFS codes can approximate a natural image.

IFS theory is an extension of classical geometry. It uses affine transformations, explained below, to express relations between parts of an image. Using only these relations, it defines and conveys intricate pictures. With IFS theory, we can describe a cloud as clearly as an architect can describe a house.

By studying the following sections, you should be able to encode and decode fascinating black-and-white image segments, such as leaf skeletons, tree shadows, spirals, and thunderheads. You should also obtain an overview of how a fully automated fractal compression system operates.

Affine transformations can be described as combinations of rotations, scalings, and translations of the coordinate axes in n-dimensional space. An example in two dimensions is

$$W(x,y) = (\tfrac{1}{2}x + \tfrac{1}{4}y + 1,\ \tfrac{1}{4}x + \tfrac{1}{2}y + 2),$$

which can also be written in matrix form as

$$W\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} .5 & .25 \\ .25 & .5 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$

This transformation moves the point (0,0) to (1,2) and moves (−1,0.5) to (0.625, 2). To confirm your understanding of the idea, you should work out where it moves the point (1,1). We denote this transformation by W; the notation $W(S)$ denotes the subimage of W on a set of points S.

Now let's see what W does to a picture of a smiling face, F, lying on the x,y plane (see figure 6). The result is a new, squeezed face $W(F)$. The affine transformation has deformed and moved the face. Notice that the eyes in the transformed face $W(F)$ are closer together than they are in F. We say that the transformation W is contractive: It always moves points closer together.

Another example of a contractive affine transformation is shown in figure 7. This time it acts on a leaf to produce a new, smaller leaf.

The general form for an affine transformation is

Figure 4: *IFS-encoded color photo of the Monterey coast (60 transformations, 100 bytes).*

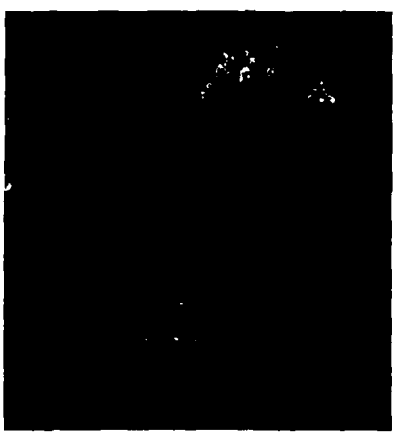


Figure 5: *IFS-encoded color image from* A Cloud Study *(30 transformations, 500 bytes).*

$$W\begin{bmatrix}x\\y\end{bmatrix} = \begin{bmatrix}a & b\\c & d\end{bmatrix}\begin{bmatrix}x\\y\end{bmatrix} + \begin{bmatrix}e\\f\end{bmatrix} = \begin{bmatrix}ax+by+e\\cx+dy+f\end{bmatrix},$$

where the coefficients a, b, c, d, e, and f are real numbers.

If we know in advance the translations, rotations, and scalings that combine to produce W, we can generate coefficient values as follows:

$$a = r\cos\theta,\ b = -s\sin\phi,$$
$$c = r\sin\theta,\ d = s\cos\phi,$$

where r is the scaling factor on x, s is the scaling factor on y, θ is the angle of rotation on x, ϕ is the angle of rotation on y, e is the translation on x, and f is the translation on y.

How can you find an affine transformation that produces a desired effect? Let's show how to find the affine transformation that takes the big leaf to the little leaf in figure 7. We wish to find the numbers a, b, c, d, e, and f for which the transformation W has the property

$$W(\text{big leaf}) \approx \text{little leaf}.$$

Begin by introducing x and y coordinate axes, as already shown in the figure. Mark three points on the big leaf (we've chosen the leaf tip, a side spike, and the point where the stem joins the leaf) and determine their coordinates (α_1,α_2), (β_1,β_2), and (γ_1,γ_2). Mark the corresponding points on the little leaf and determine their coordinates $(\tilde{\alpha}_1,\tilde{\alpha}_2)$, $(\tilde{\beta}_1,\tilde{\beta}_2)$, and $(\tilde{\gamma}_1,\tilde{\gamma}_2)$, respectively.

Determine values for the coefficients a, b, and e by solving the three linear equations

$$\alpha_1 a + \alpha_2 b + e = \tilde{\alpha}_1, \quad (1)$$
$$\beta_1 a + \beta_2 b + e = \tilde{\beta}_1, \quad (2)$$
$$\gamma_1 a + \gamma_2 b + e = \tilde{\gamma}_1, \quad (3)$$

and find c, d, and f in similar fashion from these equations:

$$\alpha_1 c + \alpha_2 d + f = \tilde{\alpha}_2, \quad (4)$$
$$\beta_1 c + \beta_2 d + f = \tilde{\beta}_2, \quad (5)$$
$$\gamma_1 c + \gamma_2 d + f = \tilde{\gamma}_2. \quad (6)$$

We recommend the use of an equation solver such as TK Solver Plus (Universal Technical Systems, Rockford, Illinois) or Eureka (Borland International, Scotts Valley, California) for finding the coefficient values. Doing it manually can be tedious.

Now that we know what a contractive

continued

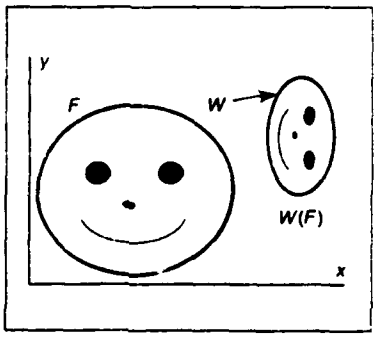


Figure 6: *An affine transformation* W *moves the smiling face* F *to a new face* W(F). *The transformation is called contractive because it moves points closer together.*

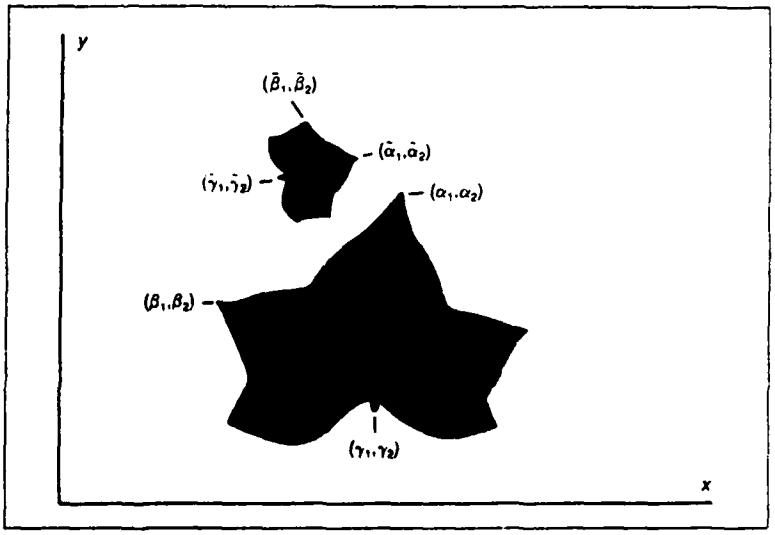


Figure 7: *Two ivy leaves fix an affine transformation* W.

Table 1: *IFS codes for a Sierpiński triangle.*

W	a	b	c	d	e	f	p
1	0.5	0	0	0.5	0	0	0.33
2	0.5	0	0	0.5	1	0	0.33
3	0.5	0	0	0.5	0.5	0.5	0.34

Table 2: *IFS codes for a square.*

W	a	b	c	d	e	f	p
1	0.5	0	0	0.5	0	0	0.25
2	0.5	0	0	0.5	0.5	0	0.25
3	0.5	0	0	0.5	0	0.5	0.25
4	0.5	0	0	0.5	0.5	0.5	0.25

Table 3: *IFS codes for a fern.*

W	a	b	c	d	e	f	p
1	0	0	0	0.16	0	0	0.01
2	0.2	-0.26	0.23	0.22	0	1.6	0.07
3	-0.15	0.28	0.26	0.24	0	0.44	0.07
4	0.85	0.04	-0.04	0.85	0	1.6	0.85

Table 4: *IFS codes for fractal tree.*

W	a	b	c	d	e	f	p
1	0	0	0	0.5	0	0	0.05
2	0.1	0	0	0.1	0	0.2	0.15
3	0.42	-0.42	0.42	0.42	0	0.2	0.4
4	0.42	0.42	-0.42	0.42	0	0.2	0.4

affine transformation is and how to find one that maps a source image onto a desired target image, we can describe an iterated function system. An IFS is a collection of contractive affine transformations. Here's an example of an IFS of three transformations:

$$W_1\begin{bmatrix}x\\y\end{bmatrix} = \begin{bmatrix}0.5 & 0.0\\0.0 & 0.5\end{bmatrix}\begin{bmatrix}x\\y\end{bmatrix} + \begin{bmatrix}0\\0\end{bmatrix},$$

$$W_2\begin{bmatrix}x\\y\end{bmatrix} = \begin{bmatrix}0.5 & 0.0\\0.0 & 0.5\end{bmatrix}\begin{bmatrix}x\\y\end{bmatrix} + \begin{bmatrix}1\\0\end{bmatrix},$$

$$W_3\begin{bmatrix}x\\y\end{bmatrix} = \begin{bmatrix}0.5 & 0.0\\0.0 & 0.5\end{bmatrix}\begin{bmatrix}x\\y\end{bmatrix} + \begin{bmatrix}.25\\.5\end{bmatrix}.$$

Each transformation must also have an associated probability, p_i, determining its "importance" relative to the other transformations. In the present case we might have p_1, p_2, and p_3. Notice that the probabilities must add up to 1. That is, $p_1 + p_2 + p_3 = 1$.

Of course, the above notation for an IFS is cumbersome. Table 1 expresses the same information in tabular form. Other examples of IFS codes are given in tables 2 through 4. Notice that an IFS can contain any number of affine transformations.

The Random Iteration Algorithm

Now let's see how to decode an arbitrary IFS code using the random iteration method. Remember that in general an IFS can contain any number, say m, of affine transformations, W_1, W_2, W_3, . . . , W_m, each with an associated probability. The following code summarizes the method:

(i) Initialize: $x=0$, $y=0$.
(ii) For $n=1$ to 2500, do steps (iii)–(vii).
(iii) Choose k to be one of the numbers 1, 2, . . . , m, with probability p_k.
(iv) Apply the transformation W_k to the point (x,y) to obtain $(\tilde{x},\tilde{y})$.
(v) Set (x,y) equal to the new point: $x=\tilde{x}$, $y=\tilde{y}$.
(vi) If $n>10$, plot (x,y).
(vii) Loop.

Applying this procedure to the transformation in table 1 produces the figure shown in figure 8—a fractal known as the Sierpiński triangle. Increasing the number of iterations n adds points to the image. Figure 9 shows the result of the random iteration algorithm applied to the data in table 3, at several stages during the process. By increasing the scale factor used in plotting, you can zoom in on the image (see figure 10). The text box on page 221 contains a BASIC implementation of the method with additional comments on programming.

You may wonder why the first 10 points are not plotted (step (vi)). This is to give the randomly dancing point time to settle down on the image. It is like a soccer ball thrown onto a field of expert players: Until someone gains control of the ball, its motion is unpredictable, or at least is independent of the players' actions. But eventually a player gets the ball, and its motion then becomes a direct result of the skill of the players. The fact that our transformation is contractive guarantees that the "ball" will eventually get to one of the "players," and that it will stay under control after that.

How do we know that the random iter-

continued

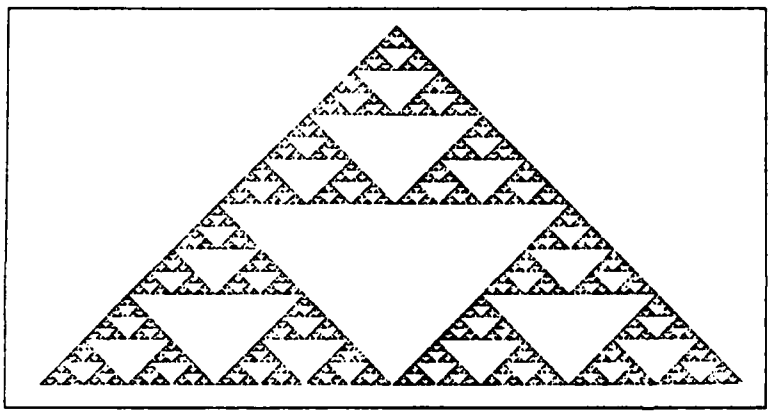


Figure 8: *The result of applying the random iteration algorithm to the IFS code in table 1. It is called the Sierpiński triangle.*

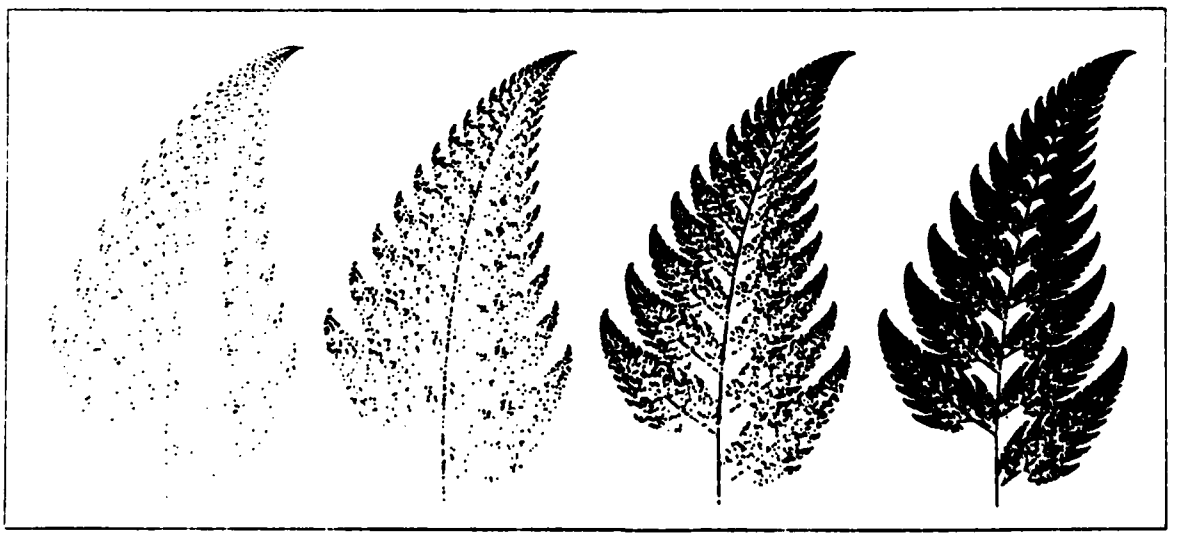


Figure 9: *A fern appears when the random iteration algorithm is applied to the IFS code in table 3.*

ation algorithm will produce the same image over and over again, independent of the particular sequence of random choices that are made? This remarkable result was first suggested by computer-graphical mathematics experiments and later given a rigorous theoretical foundation by Georgia Tech mathematician John Elton.

The Collage Theorem

Our next goal is to show a systematic method for finding the affine transformations that will produce an IFS encoding of a desired image. This is achieved with the help of the Collage Theorem.

To illustrate the method, we start from a picture of a filled-in square S in the x,y plane, with its vertices at (0,0), (1,0), (1,1), and (0,1) (see figure 11). The objective is to choose a set of contractive affine transformations, in this case W_1, W_2, W_3, W_4, so that S is approximated as well as possible by the union of the four subimages $W_1(S) \cup W_2(S) \cup W_3(S) \cup W_4(S)$. Figure 11 shows, on the left, S together with four noncovering affine transformations of it; on the right, the affine transformations have been adjusted to make the union of the transformed images cover up the square.

To find the coefficients of these transformations, we use the method described earlier in the section on iterated function systems, leading to simultaneous equations 1 through 3 and 4 through 6. The values one finds in the present case are given in table 2. When the random iteration algorithm is applied to this IFS code, the square is regenerated.

The preceding example typifies the general situation: You need to find a set of affine transformations that shrink distances and that cause the target image to be approximated by the union of the affine transformations of the image. The Collage Theorem says that the more accurately the image is described in this way, the more accurately the transformations provide an IFS encoding of it.

Figure 12 provides another illustration of the Collage Theorem. At the bottom left is shown a polygonalized leaf boundary, together with four affine transformations of that boundary. The transformed leaves taken together do not form a very good approximation of the leaf; in consequence, the corresponding IFS image (bottom right), computed using the random iteration algorithm, does not look much like the original leaf image. However, as the collage is made more accurate (upper left), the decoded image (upper right) becomes more accurate.

So, there's a fundamental stability here. You don't have to get the IFS code exactly right in order to capture a good likeness of your original image. Moreover, the IFS code is robust: Small perturbations in the code will not result in unacceptable damage to the image. In each of the above examples, we have used four transformations to encode the image. However, any number can be used.

For example, the spiral image in figure 13 can be encoded with just two contractive affine transformations. See if you can find them. Then determine the IFS transformation coefficients and input them to the random iteration algorithm to get the spiral back again.

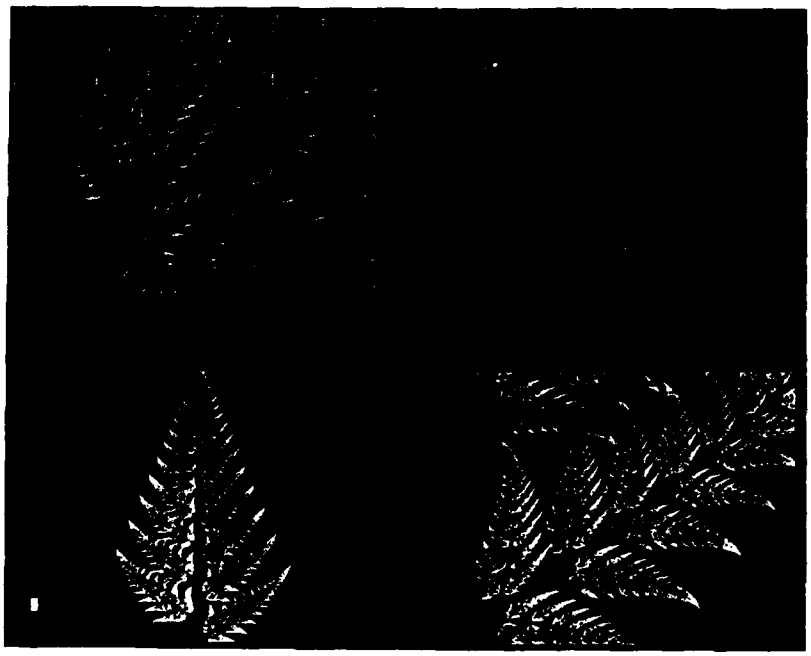


Figure 10: *Successive zooms on pieces of an IFS-encoded fern.*

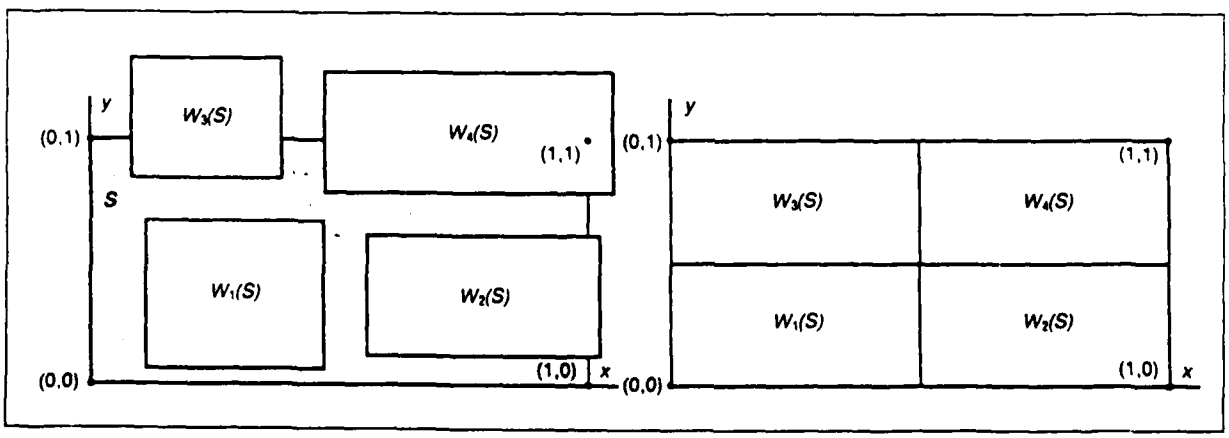


Figure 11: *The collage theorem is used to encode a classical square* S. *The correct IFS code is obtained when the four affine transformations of* S *cover* S, *as shown on the right.*

Assigning Probabilities

Once you have defined your transformations, you need to assign probabilities to them. Different choices of probabilities do not in general lead to different images, but they do affect the rate at which various regions or attributes of the image are filled in. Let the affine transformations W_i corresponding to an image I be

$$W_i \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} a_i & b_i \\ c_i & d_i \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} e_i \\ f_i \end{bmatrix},$$

where $i = 1, 2, 3, \ldots, n$. Then the amount of time that the randomly dancing point should spend in the subimage W_i is approximately equal to

$$\frac{\text{area of } W_i}{\text{area of } I}.$$

continued

IFS Decoding in BASIC

Listing A is a BASIC implementation of the random iteration algorithm. It includes the data for the Sierpiński triangle, but you can use it to process any IFS tables. In particular, you will want to try the data in tables 2, 3, and 4. Be sure to set the variable `m` correctly; it tells the program how many transformations are in the IFS.

It is also essential that the probabilities in `p()` add up to 1. For speed, the transformations should be listed in descending order of probability: the highest probability transformation first, and the lowest probability last.

The program includes variables for rescaling and translating the origin to accommodate the range of the points being plotted to the limits of your screen. If the image is too wide, decrease `xscale`; if the points are too close horizontally, increase `xscale`. Adjust `yscale` similarly to get a good vertical point spread. To move the image, adjust `xoffset` and `yoffset`.

You can do these adjustments by trial and error: Run the program; interrupt it and change the offsets and scale factors; and run it again. Or, you can replace the plot command `pset` with a command to print the values of x and y and run the program to get an exact idea of the range of points being plotted, so you can adjust the scale and offsets more precisely.

Another way to arrange the program is to have it read all the data—`m`, `a()`, `b()`, `c()`, `d()`, `e()`, `f()`, `p()`, `xscale`, `yscale`, `xoffset`, and `yoffset`—from a disk file specified by the user. Instead of reading in the coefficients a, b, c, and d, you may want to read in angles θ and ϕ and scale factors r and s, and then calculate the coefficients.

The random iteration method is computation-intensive, so we recommend use of a compiler such as Microsoft's QuickBASIC or Borland's Turbo BASIC. If your computer has a floating-point coprocessor and your compiler supports one, so much the better.

Listing A: *A BASIC program demonstrating the use of the random iteration algorithm to reconstruct an IFS-compressed image.*

```
10 'Allow for a maximum of 4 transformations in the IFS
20 DIM a(4), b(4), c(4), d(4), e(4), f(4), p(4)
30 '
40 'Transformation data, Sierpinski triangle
50 'First comes the number of transformations
60 'then the coefficients a through f and probability pk
70 'The values for pk should be in descending order.
80 DATA 3
90 DATA .5,0,0,.5,0,0,.34
100 DATA .5,0,0,.5,1,0,.33
110 DATA .5,0,0,.5,.5,.5,.33
120 '
130 'Read in the data
140 READ m
150 pt = 0  'Cumulative probability
160 FOR j = 1 TO m
170     READ a(j), b(j), c(j), d(j), e(j), f(j), pk
180     pt = pt + pk
190     p(j) = pt
200 NEXT j
210 '
220 'Set up for Graphics
230 SCREEN 3          'Select graphics screen
240 xscale = 350      'Map [0,1] onto [0,350]
250 yscale = 325      'Map [0,1] onto [0,325]
260 xoffset = 0
270 yoffset = 0       'Leave the y-origin
280 '
290 'Initialize x and y
300 x = 0
310 y = 0
320 '
330 'Do 2500 iterations
340 FOR n = 1 TO 2500
350     pk = RND
360     'The next line works for m<=4. It must be modified
370     'for values of m > 4.
380     IF pk <= p(1) THEN k = 1 ELSE IF pk <= p(2) THEN k = 2
                ELSE IF pk <= p(3) THEN k = 3 ELSE k = 4
390     newx = a(k) * x + b(k) * y + e(k)
400     newy = c(k) * x + d(k) * y + f(k)
410     x = newx
420     y = newy
430     'Use PRINT x,y instead of the PSET line
440     'to see the range of coordinates. Then fix
450     'xscale, yscale, xoffset, and yoffset
460     IF n > 10 THEN PSET (x * xscale + xoffset, y * yscale
                + yoffset)
470 NEXT n
480 '
490 LOCATE 24, 35
500 PRINT "Press any key to end.";
510 WHILE INKEY$ = ""
520 WEND
530 '
540 'Return to text screen
550 SCREEN 0
560 END
```

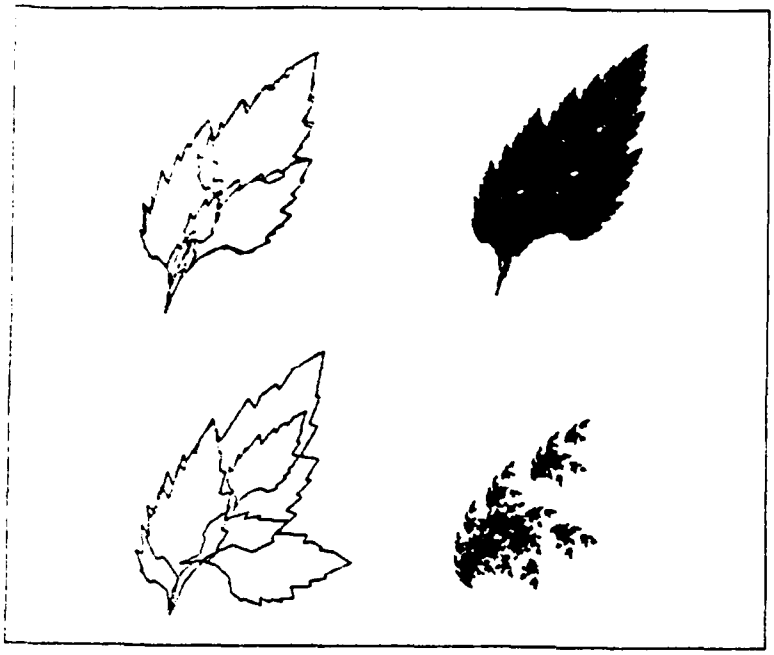


Figure 12: *The Collage Theorem is applied to a leaf. The collage at lower left isn't much good, so the corresponding IFS image, shown at lower right, is a poor approximation. But as the collage improves, upper left, so does the IFS image.*

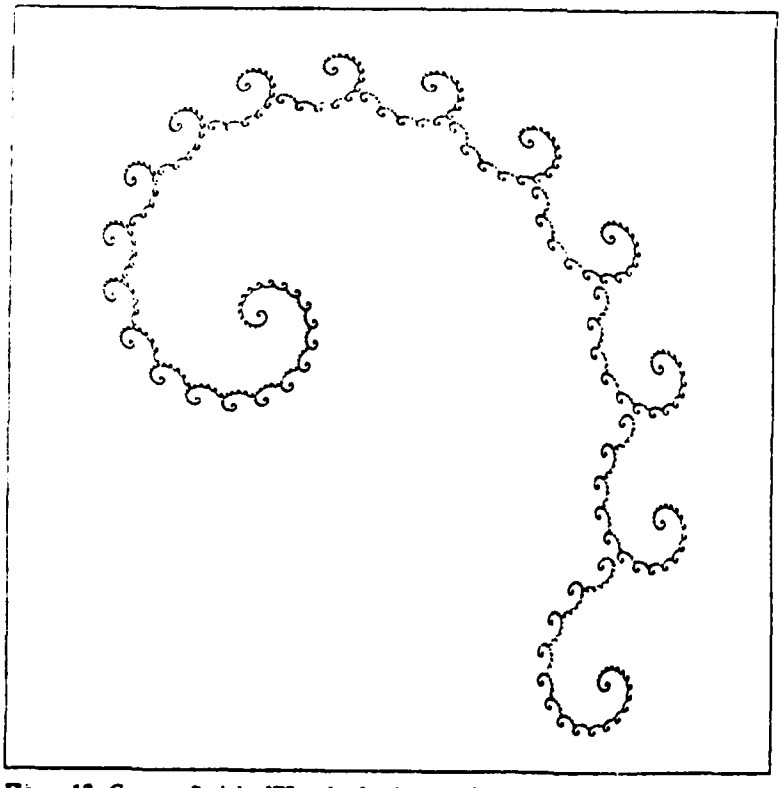


Figure 13: *Can you find the IFS codes for this spiral image? Only two transformations are needed.*

So long as $ad - cd$ is not 0, it is a standard calculus result that our ratio equals the determinant of the transformation matrix for W_i. So a good choice for the probability p_i is

$$\frac{a_i d_i - b_i c_i}{\sum_{i=1}^{n} |a_i d_i - b_i c_i|}.$$

provided none of these numbers p_i comes out to be 0. A 0 value should be replaced by a very small positive value, such as 0.001, and the other probabilities correspondingly adjusted to keep the sum of all the probabilities equal to 1.

We now summarize the compression and decompression process: An input image is broken up into segments through image-processing techniques. These image components are looked up in the IFS library using the Collage Theorem, and their IFS codes are recorded. When the image is to be reconstructed, the IFS codes are input to the random iteration algorithm. The accuracy of the reconstructed image depends only on the tolerance setting used during the collage mapping stage.

Applications

For graphics applications, we use a more sophisticated procedure that allows full-color images to be encoded. Combinatorial searching algorithms can be used to automate the collage mapping stage. Figures 2, 3, and 4 were obtained using IFS theory at compression ratios in excess of 10,000 to 1. These images were based on photographs in recent issues of *National Geographic*. A full-sequence video animation, *A Cloud Study*, was shown at SIGGRAPH '87. This was encoded at a ratio exceeding 1,000,000 to 1 and can be transmitted in encoded form at video rates over ISDN lines (ISDN stands for integrated services digital network, a concept for integrated voice and data communications). A frame from the animation is shown in figure 5.

The IFS compression technique is computation-intensive in both the encoding and decoding phases. Computations for the color images were all carried out on Masscomp 5600 workstations (dual 68020-based systems) with Aurora graphics. Complex color images require about 100 hours each to encode and 30 minutes to decode on the Masscomp.

For practical applications, you need custom hardware that can speed the encoding and decoding process. An experimental prototype, the IFSIS (iterated function system-image synthesizer), decodes at the rate of several frames per second. The IFSIS device was produced from a cooperative effort between GTRC,

DARPA, Atlantic Aerospace Electronics Corporation, and Iterated Systems, and it was demonstrated on October 5, 1987, at the third annual meeting of the Applied and Computational Mathematics Program of DARPA. It can be connected to a personal computer through a serial port; the personal computer sends the IFS codes to the device, which responds by producing complex color images on a monitor.

The IFSIS is a proof of concept for faster devices with higher resolution. Once the higher-performance IFSIS devices are combined with ISDN telecommunication, full-color animation at video rates over phone lines will be a reality.

Another area for future application of IFS encoding is automatic image analysis. What's in a picture? Does it show a spotted sandpiper or a robin? The more complex the image or the more subtle the question, the harder it becomes for an algorithmic answer to be formulated. But here's the point: Whatever the answer, it will proceed faster if stable, compressed images are used. The reason for this is that image-recognition problems involve combinatorial searching, and searching times increase factorially with the size of the image file.

During the spring of 1987, Iterated Systems was incorporated to develop commercial applications of IFS image compression. It is exciting to see how an abstract field of mathematics research is leading to new technology with implications ranging from commercial and industrial work to personal computing. ■

ACKNOWLEDGMENTS

Figures 2 through 5 were encoded by graduate students François Malassenet, Laurie Reuter, and Arnaud Jacquin. All color images were produced in the Computergraphical Mathematics Laboratory at Georgia Institute of Technology and are copyright 1987, GTRC.

BIBLIOGRAPHY

Barnsley, M. F. and S. Demko. "Iterated Function Systems and the Global Construction of Fractals." *The Proceedings of the Royal Society of London*, A399, 1985, pp. 243–275.

Barnsley, M. F., V. Ervin, D. Hardin, and J. Lancaster. "Solution of an Inverse Problem for Fractals and Other Sets." *Proceedings of the National Academy of Science*, vol. 83, April 1985.

Barnsley, M. F. *Fractals Everywhere*. Academic Press, 1988. Forthcoming.

Elton, J. "An Ergodic Theorem for Iterated Maps." *Journal of Ergodic Theory and Dynamical Systems*. Forthcoming

Mandelbrot, B. *The Fractal Geometry of Nature*. San Francisco, CA: W. H. Freeman and Co., 1982.

BYTE

THE SMALL SYSTEMS JOURNAL

January